Title

Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction.

Journal

Genome medicine, 5(3)

ISSN

1756-994X

Authors

Quon, Gerald
Haider, Syed
Deshwar, Amit G
et al.

Publication Date

2013-03-01

DOI

10.1186/gm433

Peer reviewed

Genome Medicine

METHOD Open Access

Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction

Gerald Quon[1,9], Syed Haider[2,3], Amit G Deshwar[4], Ang Cui[5], Paul C Boutros[2,6*] and Quaid Morris[1,4,7,8*]

Abstract

Tumor heterogeneity is a limiting factor in cancer treatment and in the discovery of biomarkers to personalize it. We describe a computational purification tool, ISOpure, which directly addresses the effects of variable contamination by normal tissue in clinical tumor specimens. ISOpure uses a set of tumor expression profiles and a panel of healthy tissue expression profiles to generate a purified cancer profile for each tumor sample, and an estimate of the proportion of RNA originating from cancerous cells. Applying ISOpure before identifying gene signatures leads to significant improvements in the prediction of prognosis and other clinical variables in lung and prostate cancer.

Background

Cancer patients with similar clinical and pathological characteristics can vary dramatically in their survival and response to treatment. Much of this variation is associated with differences in the molecular and cellular architecture of their tumors, suggesting that treatment decisions can be optimized based on molecular features of each individual's tumor [1]. Microarray and high-throughput sequencing technologies can profile the relative abundance of thousands of RNAs in a tumor, thereby providing a comprehensive snapshot of tumor state. These snapshots can increase the precision of patient categorizations that are traditionally based on type, size, spread, and histology [2]. Gene signatures derived from mRNA profiles have been used to identify cancer subtypes [3-5], to predict patient prognosis [6-11] and response to treatment [12,13], and to identify the site of origin [14,15]. Some of these signatures are already in routine clinical use [16-18] or undergoing trials [19].

Tumor samples drawn from patients usually exhibit significant cellular heterogeneity [2]. The proportion of healthy tissue in a sample can vary widely even among samples pre-selected to have a high cancerous cell content using pathological estimates [20-23], thereby introducing variability into expression profiles that cannot be removed by current computational pre-processing methods. This variability interferes with the development and clinical application of gene signatures by reducing the effective sample size of profiling studies, introducing confounding transcriptional signals even in moderately impure samples [24], and restricting the clinical use of gene signatures to tumor samples with sufficient cancerous cell content.

Post-operative methods for sample purification, such as laser capture micro-dissection or cell sorting, require specialized equipment, are costly, delay the diagnostic cycle, and cannot always be used. Furthermore, they may not remove all contaminating tissue, and can induce artificial cellular responses [25], while degrading samples and increasing the odds of sample confusion. A computational approach to purifying tumor profiles would address these issues.

It is possible to purify a single tumor profile computationally by representing it as a weighted average of its constituent (but hidden) cancer and non-cancerous 'normal' expression profiles, and then using statistical inference to jointly estimate both the mixture weights and the two constituent profiles. However, this is an under-determined system of equations, as there are more parameters than observations. Previous attempts to solve this problem can

* Correspondence: paul.boutros@oicr.on.ca; quaid.morris@utoronto.ca
[1]Department of Computer Science, University of Toronto, 10 King's College Road, Room 3302, Toronto, ON, Canada, M5S 3G4
[2]Informatics and Biocomputing Platform, Ontario Institute for Cancer Research, 101 College Street, Suite 800, Toronto, ON, Canada, M5G 0A3
Full list of author information is available at the end of the article

be viewed as different ways of regularizing these parameter estimates to make the problem well-determined. Most algorithms assume that multiple tumor samples in a collection are to be simultaneously purified, and that each of the tumor profiles in the collection is a mixture of a small number of shared cancer and normal profiles (that is, these algorithms constrain all normal and cancer expression profiles that constitute each tumor profile in the collection to be the same), and these methods represent each tumor profile only by their relative proportions of each shared profile [14,20,21,26-32]. Another approach assumes that the profile, $\boldsymbol{h}_n$, of the contaminating normal cells can be measured separately for all cancer patients (whose tumor profiles are indexed by n) [33,34], and thus fixes $\boldsymbol{h}_n$ to the observed value, and freely estimates the cancer profile $\boldsymbol{c}_n$ of each tumor profile n as well as the mixing proportions. The former group of methods is not amenable to downstream sample-specific analyses such as prognostic prediction because they apply too strong regularization, and estimate only a handful of patient-specific covariates. The latter group of methods requires an accurate, separate measurement of $\boldsymbol{h}_n$; these measurements are rarely available in archival datasets and are not always feasible to obtain in a clinical setting. Furthermore, the sensitivity of these methods to biological variability or measurement noise in the provided profile for $\boldsymbol{h}_n$ remains unclear and has never been tested. However, it seems likely that the sensitivity would be high, because the estimates of the cancer profiles $\boldsymbol{c}_n$ are not regularized and therefore incorporate any noise or error in the provided normal profile $\boldsymbol{h}_n$.

In this paper, we describe a new approach to computational purification, called ISOpure, which, unlike previous approaches, is able to estimate a distinct cancer profile for each tumor sample; this cancer profile is robust to noise and does not require a matching normal profile. Using a dataset of 834 lung and prostate tumors, we found that ISOpure reduces inter-tumor variability caused by non-cancerous tissue contamination, leading to a significant increase in the power and accuracy of clinical prediction models. Using ISOpure to preprocess non-small cell lung adenocarcinoma expression profiles, we produced a validated gene signature that is a statistically significant predictor of prognosis for all lung adenocarcinoma tumors and for stage I tumors only.

Methods

The challenge of computational purification

The challenge of computational purification is to decompose each tumor profile $\boldsymbol{t}_n$ (a vector of length G) into its component cancer profile (the vector $\boldsymbol{c}_n$), and normal profile (the vector $\boldsymbol{h}_n$), and estimate a scalar, α_n, that represents the fraction of the tumor sample RNA that was contributed by cancer cells. This estimation is typically made using a procedure that sets the parameters ($\boldsymbol{c}_n$, $\boldsymbol{h}_n$, and α_n) in order to minimize the reconstruction error, represented here by the vector $\boldsymbol{e}_n$:

$$\boldsymbol{t}_n = \alpha_n \boldsymbol{c}_n + (1 - \alpha_n)\, \boldsymbol{h}_n + \boldsymbol{e}_n \quad (1)$$

Without further constraints on the parameters, equation (1) is an ill-defined problem because there are $2G+1$ parameters to estimate ($\boldsymbol{c}_n$, $\boldsymbol{h}_n$, and α_n) but only G observations ($\boldsymbol{t}_n$) so there is a continuum of solutions that satisfy equation (1) with zero error, suggesting that these solutions are over-fitting the problem. Computational purification methods apply different 'hard' and 'soft' constraints (also known as regularizations) to the parameters to ensure a unique, interpretable solution. Regularization strategies score the parameters based on how well they reflect prior assumptions about their likely values. For example, ISOpure assumes that the vector $\boldsymbol{h}_n$ is similar to one or more profiles of normal tissue that are input into the algorithm. Because the choice of regularization determines the solution to equation (1), the success of a computational purification method depends on the suitability of the regularizations that it applies. In the Results section, we evaluate ISOpure and other regularization strategies based on how much they improve prognostic models applied to the tumor profiles and how well they reproduce pathological evaluations of the tumors. Typically, good regularization strategies introduce a sufficiently strong bias into parameter estimation that they favor solutions to equation (1) that have non-zero error and therefore avoid over-fitting. Thus, assumptions about the distribution of $\boldsymbol{e}_n$ also influence the solution to equation (1). Different assumptions lead to different objective functions in the estimation, and can lead to different optimization procedures.

The following sections describe our ISOpure method in detail. In brief, ISOpure is based on a statistical model that represents the tumor profile as a sample from a multinomial distribution. The multinomial distribution is parameterized by a discrete probability distribution (represented by the vector $\hat{\boldsymbol{x}}_n$) that ISOpure attempts to decompose into the cancer profile $\boldsymbol{c}_n$ and the normal profile $\boldsymbol{h}_n$. The reconstruction error $\boldsymbol{e}_n$ from equation (1) can be interpreted as sampling noise from the multinomial distribution, but it is not explicitly represented in the ISOpure model. ISOpure makes two prior assumptions to avoid over-fitting: it assumes that $\boldsymbol{h}_n$ is a convex combination of the normal profiles provided to the algorithm, and that the cancer profiles $\boldsymbol{c}_1$, $\boldsymbol{c}_2$,..., $\boldsymbol{c}_N$ in the cohort are clustered together around a 'reference cancer profile', $\boldsymbol{m}$, which is also inferred from the data. The parameters of the ISOpure model, which include the individual cancer and normal profiles and the reference cancer profile, are fit by maximum *a posteriori* (MAP) estimation in the

statistical model (equations 3 to 9) that encodes the ISOpure assumptions.

ISOpure overview

Below we provide a brief overview of the major features of the algorithm. The following sections contain a description of the parameters of the ISOpure model, a formal specification of this statistical model (equations 3 to 9) and a step-by-step guide to the inference procedure that ISOpure uses to fit its model. Our notation is as follows: lower case letters (for example, α_n) represent scalar parameters or indices, bold lowercase letters (for example, $\boldsymbol{c}_n$) represent column vectors (which could be inputs or parameters), capital letters (for example, G) represent scalar constants, and bold capital letters (for example, $\boldsymbol{B}$) represent matrices.

ISOpure inputs

In our comparisons, the following input data were assumed to be available to ISOpure and other algorithms:

a) $\boldsymbol{t}_1, \boldsymbol{t}_2,..., \boldsymbol{t}_N$ is a set of N tumor profiles. Each profile is represented by a vector of G non-negative (that is, 0 or greater) elements, where each element represents the measured expression level of a transcript. G is typically on the order of 10,000. Microarray intensities should be normalized but not log-transformed before input into ISOpure, as the algorithm interprets each element as a normalized count of the number of copies of each transcript present in the sample.

b) $\boldsymbol{b}_1, \boldsymbol{b}_2,..., \boldsymbol{b}_R$ is a set of R healthy profiles, defined as above, that are ideally collected using the same protocol as was used to collect the tumor profiles. Note that we expect R to be less than N, and we do not require that any of the healthy profiles be matched to a tumor.

ISOpure outputs

ISOpure estimates the following variables from the input data:

a) $\boldsymbol{c}_1, \boldsymbol{c}_2,..., \boldsymbol{c}_N$ is a set of cancer profiles, each of length G, that represent the tumor profiles purified of normal contamination. Cancer profile $\boldsymbol{c}_n$ corresponds to input tumor profile $\boldsymbol{t}_n$, where element g of the vector $\boldsymbol{c}_n$ ($c_{n,g}$) represents the estimate of the relative abundance of transcript g in the cancer cell population of tumor n. In ISOpure, each vector $\boldsymbol{c}_n$ can be interpreted as a probability distribution over the transcripts; in other words, the elements of $\boldsymbol{c}_n$ are non-negative and sum to one, and $c_{n,g}$ represents the probability of picking transcript g if a random sample is taken from the population of transcripts in the cancerous cell population in tumor n.

b) $\alpha_1, \alpha_2,..., \alpha_N$ is a set of 'tumor purity' estimates, where α_n is the estimated fraction of RNA in tumor sample n that was contributed by the cancer cells. It can be interpreted as an estimate of the probability that a random transcript from the n^{th} (mixed) tumor cell population (represented by $\boldsymbol{t}_n$) originated from a cancerous cell.

Summary of key features of ISOpure

ISOpure employs two main regularization strategies. First, ISOpure assumes that each normal (that is, healthy) profile $\boldsymbol{h}_n$ can be represented by a weighted combination of the available healthy tissue profiles $\boldsymbol{b}_1, \boldsymbol{b}_2,..., \boldsymbol{b}_R$. In other words, ISOpure replaces equation (1) with

$$\boldsymbol{t}_n = \alpha_n \boldsymbol{c}_n + \sum_{r=1}^{R} \theta_{n,r} \boldsymbol{b}_r + \boldsymbol{e}_n, \quad (2)$$

where $\theta_{n,1}, \theta_{n,2},..., \theta_{n,R}$ are parameters fit by ISOpure. It further requires that these new parameters are non-negative and that

$$\alpha_n + \sum_{r=1}^{R} \theta_{n,r} = 1.$$

Thus, $\theta_{n,r}$ can be interpreted as the proportion of the transcripts in the n^{th} tumor arising from the 'tissue' represented by profile $\boldsymbol{b}_r$. (Note that to simplify notation, we occasionally indicate $\theta_{n,1}, \theta_{n,2},..., \theta_{n,R}$ and α_n using the vector $\boldsymbol{\theta}_n$ of length $R+1$ whose r^{th} element is $\theta_{n,r}$ for $r < R+1$ and whose $R+1^{\text{st}}$ element is set to α_n.) This regularization assumption reduces the number of output parameters to $R+G$ ($\alpha_n, \theta_{n,1}, \theta_{n,2},..., \theta_{n,R-1}, \boldsymbol{c}_n$). In our experiments, R was at most 50. Although the number of estimated parameters is still greater than the number of observations (G) and, therefore, there still remains a continuum of solutions to equation (2) that have zero error, the large reduction in parameter number allows us to apply a weaker regularization to $\boldsymbol{c}_n$ and still avoid over-fitting. This strategy also ensures that the contaminating normal profile $\boldsymbol{h}_n$, implicitly estimated by the algorithm, is similar to the normal tissue types represented by the input profiles $\boldsymbol{b}_r$ to the algorithm. The other regularization strategy used in ISOpure is that it favors solutions in which the values of $\boldsymbol{c}_1, \boldsymbol{c}_2,..., \boldsymbol{c}_N$ are clustered together. It encodes this using a scoring function that encourages $\boldsymbol{c}_n$ to be similar to an estimated 'reference cancer profile', $\boldsymbol{m}$. In other words, ISOpure assumes that the tumor samples in the same collection have similar expression profiles except for some sample-specific deviations that influence prognosis and response to therapy; this assumption is more accurate when the tumors are of the same subtype (for example, adenocarcinomas of the lung [1]). The vector $\boldsymbol{m}$ is a parameter of the algorithm that is estimated from the tumor profile data, and itself has a regularization applied to it to bias its estimate toward values that are close to the normal profiles. This modeling choice reflects an assumption that in general, profiles of cancerous tissue are similar (but not identical) to those of

the tissue of origin of the tumor type. We have previously reported [14] that this assumption improves the accuracy of tumor purity estimation.

Additional estimated parameters of ISOpure

Our regularization strategy incorporates the Dirichlet probability density function into our scoring functions. This choice allows us to use the statistical inference method described below to estimate the parameter values. The Dirichlet distribution is a continuous multivariate distribution over discrete probability distributions (that is, vectors of pre-determined size that contain non-negative elements that sum to one). We use the Dirichlet for both $\boldsymbol{\theta}_n$ and $\boldsymbol{c}_n$ because they are both discrete probability distributions. The probability density function associated with the Dirichlet has two parameters (termed hyper-parameters because they are the parameters of distributions over model parameters): a mean vector (which determines the mean of the Dirichlet distribution) and a scalar strength parameter that controls how quickly the score decreases from the mode of the Dirichlet distribution. We also estimate the following hyper-parameters from the tumor data: $\boldsymbol{\nu}$, k_n (for n = 1 to N), k', and $\boldsymbol{\omega}$. These additional parameters are formally defined below in the statistical model provided in equations (3 to 9), but in brief $\boldsymbol{\nu}$ represents both the mean and strength of a Dirichlet distribution over $\boldsymbol{\theta}_n$; k_n represents the strength parameter of the Dirichlet distribution over $\boldsymbol{c}_n$ given $\boldsymbol{m}$; k' represents the strength parameter of the Dirichlet distribution over $\boldsymbol{m}$; $\boldsymbol{\omega}$ represents the weights on the normal profiles $\boldsymbol{b}_r$ used to make the weighted combination that forms the mean parameter vector for the Dirichlet distribution over $\boldsymbol{m}$.

ISOpure algorithm

As mentioned above, the parameter estimates that achieve the best score for a given regularization strategy typically yield non-zero error (represented by $\boldsymbol{e}_n$) in reconstructing the tumor profile in equation (2). Regularization strategies therefore must also determine the optimal trade-off between decreasing the error $\boldsymbol{e}_n$ in equation (2) and improving the scores of the parameters (that is, $\boldsymbol{c}_n$ and $\boldsymbol{\theta}_n$ for all values of n) under the Dirichlet distributions encoding our prior assumptions. We formalize the minimization of error $\boldsymbol{e}_n$ as the maximization of the probability of a count vector $\boldsymbol{x}_n$ (derived by discretization of $\boldsymbol{t}_n$) under a multinomial distribution whose probability vector over transcripts is

$$\hat{\boldsymbol{x}}_n = \alpha_n \boldsymbol{c}_n + \sum_{r=1}^{R} \theta_{n,r} \boldsymbol{b}_r$$

(that is, $\hat{\boldsymbol{x}}_n$ is a normalized reconstruction of the tumor profile $\boldsymbol{x}_n$ based on the model parameters). The score of a given parameter setting is the product of the score of the parameters under the Dirichlet prior distributions and the probability of the discretized tumor profiles under the multinomial distribution defined by $\hat{\boldsymbol{x}}_n$. Maximizing this score is equivalent to MAP estimation under the statistical model described below. Note that we estimate the parameters of all of the tumor profiles simultaneously because some of the model parameters (for example, $\boldsymbol{m}$) depend on all tumor profiles.

ISOpure pre-processing and data transformation

In this step, ISOpure applies some simple transforms to the inputs in order to place them in an appropriate form for the model.

The first transform is to discretize the tumor profiles $\boldsymbol{t}_n$ by rounding each element to the nearest non-negative integer to make the count vector $\boldsymbol{x}_n$. The statistical model underlying ISOpure interprets the elements of the tumor profile as a count of the number of transcripts of that type (gene or transcript isoform) observed in the sample. Ideally, the tumor profiles should be rescaled so that the total number of observations (that is, the sum of the elements) after discretization is approximately the same across all tumor profiles, in order to balance the influence that each tumor profile has on the shared parameters. In the tumor profiles we used in our experiments, the sum of the elements in each of the discretized profiles after robust multi-array average (RMA) normalization was on the order of 10^7. Profiles may need to be rescaled before discretization if their sum is much less than this, to ensure adequate precision in the discretization.

The second transform is to divide each normal profile $\boldsymbol{b}_r$ by the sum of its elements. After this transformation, each profile $\boldsymbol{b}_r$ sums to one, allowing them to be interpreted as a discrete probability distribution over transcripts.

ISOpure statistical model

The full ISOpure model is defined as follows (the probability density and mass functions of the Dirichlet and Multinomial distributions, respectively, are given in Table 1).

$$\boldsymbol{B} = \left[\boldsymbol{b}_1 \cdots \boldsymbol{b}_R\right] \tag{3}$$

$$\hat{\boldsymbol{x}}_n = [\boldsymbol{B} \quad \boldsymbol{c}_n]\,\boldsymbol{\theta}_n \tag{4}$$

$$p(\boldsymbol{\theta}_n|\boldsymbol{\nu}) = \text{Dirichlet}(\boldsymbol{\theta}_n|\boldsymbol{\nu}) \tag{5}$$

$$p(\boldsymbol{x}_n|\boldsymbol{B}, \boldsymbol{\theta}_n, \boldsymbol{c}_n) = \text{Multinomial}\left(\boldsymbol{x}_n|\hat{\boldsymbol{x}}_n\right) \tag{6}$$

$$p(\boldsymbol{c}_n|k_n, \boldsymbol{m}) = \text{Dirichlet}(\boldsymbol{c}_n|k_n\boldsymbol{m}) \tag{7}$$

$$p\left(\boldsymbol{m}|k', \boldsymbol{B}, \boldsymbol{\omega}\right) = \text{Dirichlet}\left(\boldsymbol{m}|k'\boldsymbol{B}\boldsymbol{\omega}\right) \tag{8}$$

Table 1 Probability density and mass functions of probability distributions used to define ISOpure.

Probability density/mass function[a, b]
$\text{Dirchlet}(\boldsymbol{x}\|\boldsymbol{a}) = \frac{\Gamma\left(\sum_{k=1}^{K} a_k\right)}{\prod_{k=1}^{K}\Gamma(a_k)} \prod_{k=1}^{K} x_k^{a_k-1}$
$\text{Multinomial}(\boldsymbol{y}\|\boldsymbol{\pi}) = \frac{\left(\sum_{k=1}^{K} y_k\right)!}{\prod_{k=1}^{K} y_k!} \prod_{k=1}^{K} \pi_k^{y_k}$

[a]The parameters *x*, *a*, *y*, and π are all assumed to be vectors of length *K*. Note that the canonical parameter of the multinomial distribution that represents the number of trials is implicit here (that is, the number of trials here is defined as the sum of all elements of *y*).
[b]$\Gamma(a)$ is the gamma function of scalar a.

where $[\boldsymbol{v}\ \boldsymbol{w}]$ indicates the matrix formed by horizontally appending column vectors and/or matrices $\boldsymbol{v}$ and $\boldsymbol{w}$ and $\boldsymbol{M}\boldsymbol{v}$ indicates a matrix-vector product of $\boldsymbol{M}$ and $\boldsymbol{v}$. We estimate the parameters $\boldsymbol{\theta}_n$ (including α_n), $\boldsymbol{c}_n$, $\boldsymbol{v}$, $\boldsymbol{m}$, k_n, k', and $\boldsymbol{\omega}$ using a two-step approach to maximize the complete likelihood function of this model:

$$\mathbb{L} = p(\boldsymbol{m}|k', \boldsymbol{B}, \boldsymbol{\omega}) \prod_{n=1}^{N} p(\boldsymbol{c}_n|k_n, \boldsymbol{m}) \cdot p(\boldsymbol{\theta}_n|\boldsymbol{v}) \cdot p(\boldsymbol{x}_n|\boldsymbol{B}, \boldsymbol{\theta}_n, \boldsymbol{c}_n) \quad (9)$$

Figure 1 shows a geometric interpretation of these two steps. The corresponding probabilistic graphical model illustrating equations (3 to 9) is shown in Figure 2. The first step of ISOpure is similar to previous deconvolution algorithms, and is nearly identical to the ISOLATE

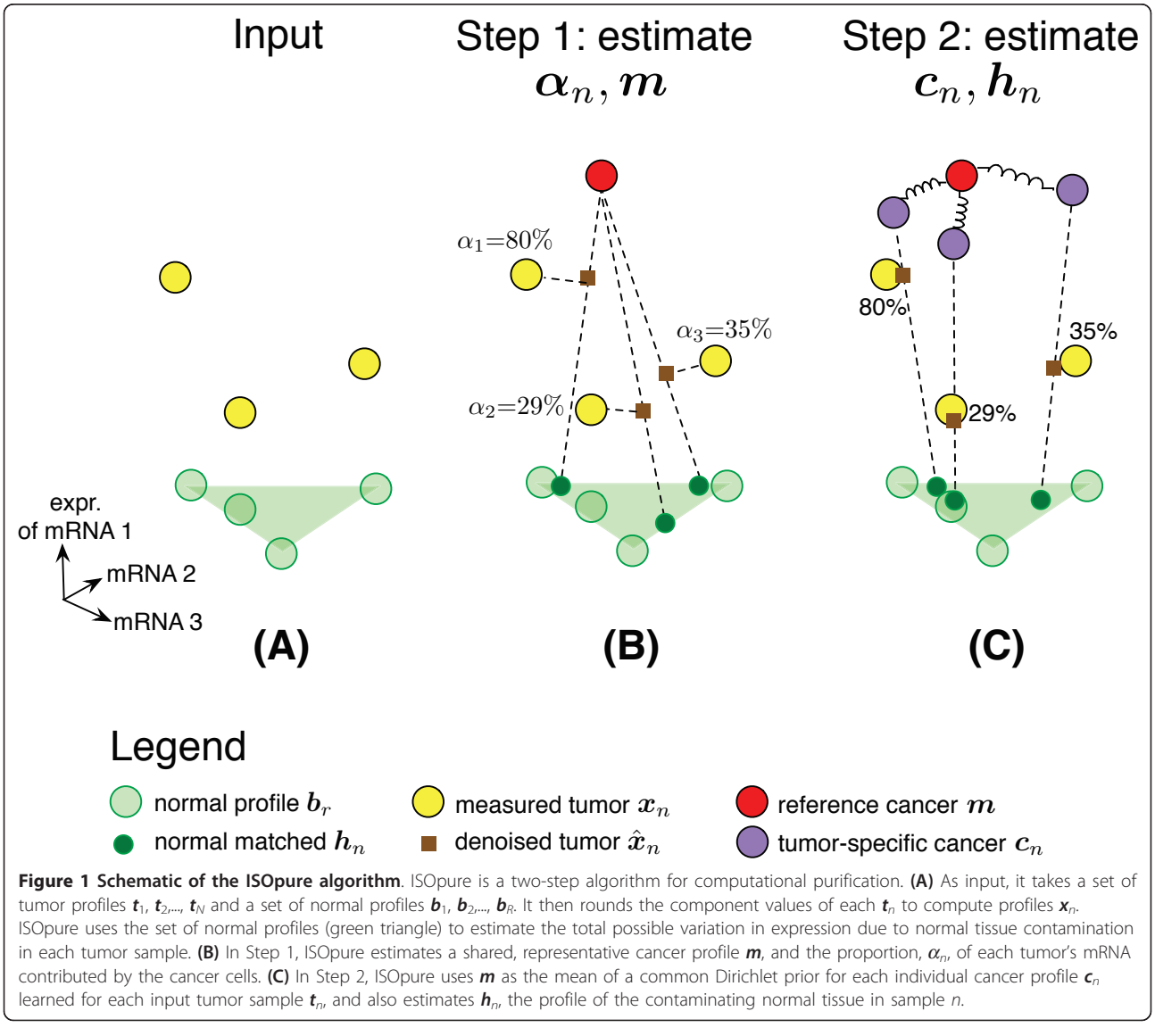

Figure 1 Schematic of the ISOpure algorithm. ISOpure is a two-step algorithm for computational purification. **(A)** As input, it takes a set of tumor profiles $\boldsymbol{t}_1$, $\boldsymbol{t}_2$,..., $\boldsymbol{t}_N$ and a set of normal profiles $\boldsymbol{b}_1$, $\boldsymbol{b}_2$,..., $\boldsymbol{b}_R$. It then rounds the component values of each $\boldsymbol{t}_n$ to compute profiles $\boldsymbol{x}_n$. ISOpure uses the set of normal profiles (green triangle) to estimate the total possible variation in expression due to normal tissue contamination in each tumor sample. **(B)** In Step 1, ISOpure estimates a shared, representative cancer profile $\boldsymbol{m}$, and the proportion, α_n, of each tumor's mRNA contributed by the cancer cells. **(C)** In Step 2, ISOpure uses $\boldsymbol{m}$ as the mean of a common Dirichlet prior for each individual cancer profile $\boldsymbol{c}_n$ learned for each input tumor sample $\boldsymbol{t}_n$, and also estimates $\boldsymbol{h}_n$, the profile of the contaminating normal tissue in sample *n*.

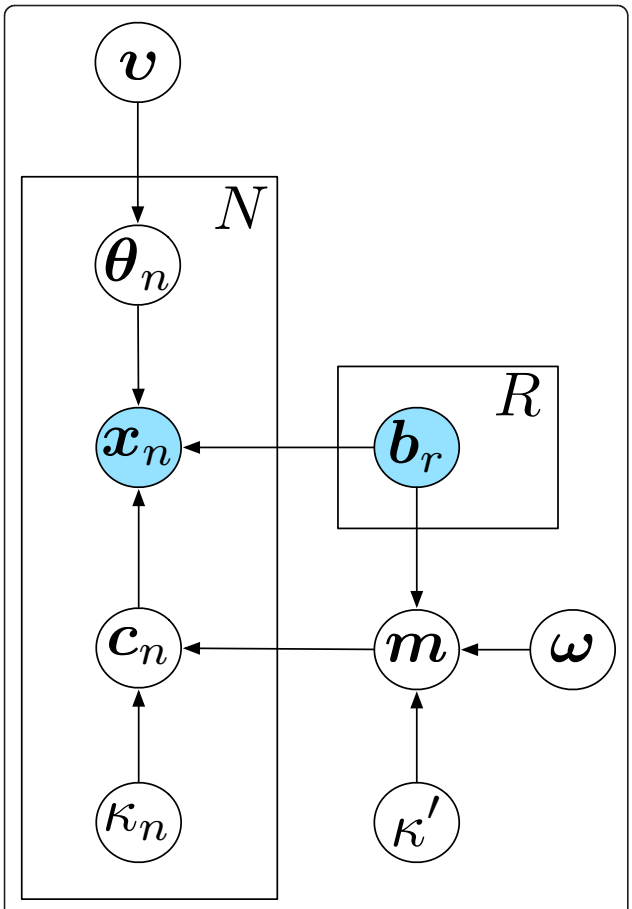

Figure 2 Graphical model representation of the ISOpure model. This directed graph represents the joint probability distribution over the input data (blue circles) and the estimated model parameters (white circles), conditioned on the estimated model hyper-parameters (white circles), in the ISOpure model (equations 3 to 9). The conditional probability of each variable depends only on its parents in the graph. To avoid explicitly representing all N tumor profiles $\boldsymbol{c}_n$ and R normal profiles $\boldsymbol{b}_r$ in the graph (and other associated variables), we used 'plate' notation by drawing one representative node per variable, and enclosing these variables in a plate (rectangular box), where the total number of instances of each enclosed variable is indicated by the fixed constant in the upper right corner of the plate.

(Identification of Sites of Origin by Latent Variables) [14] model we developed previously. The second step of the algorithm is novel. The two regularization strategies of ISOpure greatly reduce the effective number of parameters to be estimated, thereby transforming tumor-specific computational purification into a statistically well-defined estimation problem.

ISOpure Step 1: Estimate tumor purities $\alpha_1,\alpha_2,...,\alpha_N$ and the reference cancer profile m using the collection of tumor profiles

In this step, ISOpure performs MAP estimation in the statistical model. MAP estimation here is the numerical optimization of a 'complete likelihood' function that is determined based on the underlying statistical model and defined by equation (9). To simplify the optimization and focus on estimating $\alpha_1,\alpha_2,...,\alpha_N$ and $\boldsymbol{m}$, we initially force all $\boldsymbol{c}_n$ values to be exactly equal to $\boldsymbol{m}$ (that is, we set $k_n = \infty$ for all n in Step 1). Because all cancer profiles $\boldsymbol{c}_n$ are forced to equal $\boldsymbol{m}$ throughout Step 1, the estimation of $\boldsymbol{m}$ tries to simultaneously minimize the reconstruction error $\boldsymbol{e}_n$ (from equation (2)) of all tumors $\boldsymbol{t}_n$, and therefore the estimates of α_n depend on one another and they are optimized as a group. Note that to perform this estimation we must also estimate $\theta_{n,1}$, $\theta_{n,2},...,\theta_{n,R}$, although we re-estimate these values in Step 2. We also estimate the hyper-parameters (k', $\boldsymbol{\omega}$, and $\boldsymbol{\upsilon}$) that specify the Dirichlet distributions over $\boldsymbol{m}$ and $\theta_{n,1}$, $\theta_{n,2},...,\theta_{n,R}$ and α_n. When doing this, we require that $k' \geq 1/min_{r,g}\ b_{r,g}$ (where $b_{r,g}$ indicates the g^{th} element of $\boldsymbol{b}_r$) to ensure that the corresponding Dirichlet density function does not assign infinite density in the limit of one of the elements of $\boldsymbol{m}$ going towards zero. To estimate our parameters in Step 1 (and Step 2), we run 35 iterations of an optimization procedure that maximizes the complete likelihood function via block coordinate descent from a randomized starting point (see the ISOpure implementation in Additional File 1). Each iteration of this optimization procedure uses the Polak-Ribière conjugate gradient descent method [35] to estimate variables of the same type simultaneously (where we assign the same letter to variables of the same type in equations (3 to 9)) and cycles through each variable type once per iteration. We found that 35 iterations of this optimization procedure yielded a relative change in log likelihood of less than 10^{-8} between the final two iterations. To find a good local (and possibly global) maximum, we use multiple random initializations (10 in our experiments) and take the one that achieves the highest complete likelihood.

ISOpure Step 2: estimate individual cancer profiles c_n for each tumor profile

In this step, we fix α_1, $\alpha_2,...,\alpha_N$ and $\boldsymbol{m}$ to the values estimated in Step 1, and use MAP estimation to estimate the tumor-specific cancer profiles $\boldsymbol{c}_1$, $\boldsymbol{c}_2,...,\boldsymbol{c}_N$ output by the model, and to re-estimate $\theta_{n,1}$, $\theta_{n,2},...,\theta_{n,R}$ (for all n). We also estimate the hyper-parameters k_n and re-estimate $\boldsymbol{\upsilon}$; these hyper-parameters specify the Dirichlet distributions over $\boldsymbol{c}_n$ and $\theta_{n,1}$, $\theta_{n,2},...,\theta_{n,R}$ (as described above). Similar to Step 1, we require that $k_n \geq 1/min_g\ m_g$ for all n. The complete likelihood function for this step is in equation (9), and is optimized using the same algorithm as in Step 1.

ISOpure post-processing and data transformation

The main output of the ISOpure implementation in Additional File 1 are the tumor purity estimates α_n and

the cancer profiles $\boldsymbol{c}_n$. To put the estimated profiles on the same scale as the original tumor profiles, we multiply $\boldsymbol{c}_n$ by $S_n = \sum_{g=1}^{G} t_{n,g}$.

ISOpure summary

The novel contribution of ISOpure over models such as ISOLATE is the ability to estimate per-tumor cancer profiles. ISOpure is freely available as a MATLAB package (Additional File 1) that can also be run in the open-source Octave environment with a small number of modifications; the latest version is maintained here [36]. All ISOpure cancer profiles are available online (see Additional File 2; see Additional File 3; see Additional File 4; see Additional File 5).

The ISOpure-evenprior algorithm

We hypothesized that the key feature of ISOpure that enables accurate deconvolution is the assumption that the individual cancer profiles $\boldsymbol{c}_n$ are clustered, as encoded by the Dirichlet prior in equation (7). To test this hypothesis, we designed another model, ISOpure-evenprior, that is exactly the same as ISOpure (equations (3 to 9)) except that it replaces the prior defined in equation (7) with one where all the components of $\boldsymbol{m}$ are replaced with $1/G$ as follows, where $\mathbf{1}$ is a column vector of ones of length G:

$$p(\boldsymbol{c}_n|k_n, \boldsymbol{m}) = \text{Dirichlet}\left(\boldsymbol{c}_n|k_n\tfrac{1}{G}\mathbf{1}\right) \tag{10}$$

Application of the Clarke method for computational purification

To benchmark the prognostic performance of ISOpure against existing methods, we considered the Clarke [33] and Gosink [34] methods because they are the only existing methods that can be used to estimate per-tumor cancer profiles. Because the Clarke method is designed to be a robust version of the Gosink method, we tested only the Clarke method.

We downloaded the source code for the method from the web site of the authors of this method [37]. We modified the code to implement the knee-finding algorithm as presented in the original paper, as the available code did not implement it, and confirmed using the provided data that the knee-finding algorithm reproduced the same results as the original work. Because none of the tumor datasets processed in this study included matched normal profiles for each tumor, as required by the Clarke method, we used Spearman rank correlation to identify the most similar normal profile $\boldsymbol{b}_r$ for each input tumor sample, and used that normal profile as the matched normal for input into the method. Finally, because the provided code only estimates the tumor purity α_n for each tumor sample n, we estimated a tumor-specific cancer profile $\boldsymbol{c}_n$ as suggested by the authors as follows

$$\boldsymbol{c}_n = \frac{\boldsymbol{t}_n - (1-\alpha_n)\,\boldsymbol{b}_{f(n)}}{\alpha_n} \tag{11}$$

where $f(n)$ is the index of the selected matched normal. For our implementation of this algorithm, see Additional File 6.

In our prediction benchmarks, we evaluated the Clarke-based cancer profiles using exactly the same procedure we used to evaluate the ISOpure-estimated cancer profiles, as outlined below in the 'Gene signature identification and testing' section.

Predicting prognosis using the matrix factorization method

We also tested a matrix factorization-based approach to determine whether mixture proportions estimated by deconvolution algorithms (that cannot estimate individual cancer profiles [14,20,21,26-32]) could still be useful for prognostic prediction. In our prediction benchmarks, we concatenated the mixture proportions estimated by Step 1 of ISOpure for each tumor profile n (parameters $\theta_{n,1}$, $\theta_{n,2}$,..., $\theta_{n,R}$, and α_n) into a 'mixture proportion profile' vector, then evaluated these mixture proportion profiles for predictive performance in the same manner as we evaluated the ISOpure cancer profiles, as described in the 'Gene signature identification and testing' section below.

Array data processing

Raw data from the Bhattacharjee study [22] were downloaded in the form of CEL files. These data were pre-processed using the RMA algorithm [38] implemented in the affy package (version 1.22.1) for the R statistical environment (version 2.9.2). Updated ProbeSet mappings to Entrez Gene IDs were used [39] (hgu95av2hsentrezgcdf, version 12.0.0), and only adenocarcinomas of the lung were considered in this study (the predominant histological subtype, and the same subtype represented in all other patient cohorts used here). This dataset included 17 healthy samples, of which 14 were used for purification with ISOpure and three were treated as blind control samples (NL1179, NL1675, and NL1698). Of the 127 lung adenocarcinoma samples, 32 were annotated with tumor cellularity estimates made by two pathologists in the original dataset. For evaluation of ISOpure estimates of tumor purity, we removed 12 samples for which the pathologists' estimates differed by more than one standard deviation (SD) of the differences in their estimates (13.7%), leaving 20 samples for analysis. We did this because the two pathologists differed by as much as 50%

in their estimates of cancerous tissue content (see Additional File 7: Figure S1).

Raw data from the Beer [40] study was processed using the same pipeline as the Bhattacharjee study, except that we used ProbeSet mappings to Entrez Gene IDs appropriate to the specific platform used in that study (hu6800hsentrezgcdf version 12.0.0). This dataset comprised 86 tumor samples and 10 healthy samples that were used for purification with ISOpure.

Raw data from each of the four cohorts of the Director's Challenge dataset [41] were separately co-normalized using the RMA algorithm with 49 healthy lung samples profiled on the same platform by Landi and colleagues [42], using the affy package (version 1.24.2) for the R statistical environment (version 2.10.1). Again, updated ProbeSet mappings to Entrez Gene IDs were used (hgu133ahsentrezgcdf version 12.1.0). Associated survival data were downloaded from online supplementary files. From the original set of 443 patients, three patients were removed from the prognostic prediction analysis because of missing data for survival time (NCI_-lung216_U133A) or stage (Moff-0683H and Moff-0928E).

Raw data from the Wang study [21] were normalized using the RMA algorithm implemented in the affy package (version 1.26.1) for the R statistical environment (version 2.11.0). To map probe names to Entrez Gene IDs, an updated CDF file was used (hgu133plus2hsentrezgcdf version 13.0.0). After normalization, ISOpure was run on the 109 prostate tumor expression profiles, using 32 biopsy and 13 autopsy samples as the healthy tissue panel (both of which are reported as cancer-free).

Raw data from the Wallace study [43] were normalized using the RMA algorithm implemented in the affy package (version 1.26.1) for the R statistical environment (version 2.11.0). To map probe names to Entrez Gene IDs, an updated CDF file was used (hgu133a2hsentrezgcdf version 13.0.0). This dataset included 69 prostate tumor profiles and 20 healthy profiles, though two healthy profiles were removed because they were collected using pooled RNA samples.

Gene signature identification and testing

Figure 3 outlines our overall strategy for gene signature identification and testing for the prediction of prognosis of lung cancer patients. In brief, for each benchmark, we trained two elastic net-regularized Cox proportional hazards (CPH) models on the tumor gene expression profiles and the associated survival data. One of these models was trained on unpurified profiles and the other on profiles purified with ISOpure. We use the term 'gene signature' to refer to a learned CPH model that consists of a list of genes and their non-zero regression coefficients. Note that the elastic net regularization applies only to the CPH model, and is separate from the regularization used in the purification procedure. For the lung adenocarcinoma section, we ran two different benchmarks. In the first, we grouped the four independent cohorts of the Director's Challenge dataset into a training dataset of 254 patients (comprised of the HLM and MI cohorts) and testing dataset of 186 patients (comprised of the DFCI and MSKCC cohorts) as previously described [41], using either the ISOpure cancer profiles or the original unpurified profiles. We first median-centered and unit-normalized all gene expression measurements in the training dataset to bring the profiles to the same scale, as previously described [9]. Then, using the glmnet package (version 1.3) [44] on an installation of R (version 2.11.0), we ran five-fold cross-validation using the glmnetCV command with default parameters and $\alpha = 0.1$ to identify the specific gene signature that maximized cross-validation performance on the training dataset. The identified gene signature was re-run on the entire training dataset to find its median risk score, to be used as a threshold for identifying low-risk and high-risk patients in the testing dataset. We then median-centered and unit-normalized all gene expression measurements in the testing dataset, and used the identified gene signature to independently assess risk scores for each of the 186 patients in the testing dataset. These risk scores were used to categorize patients into low-risk or high-risk groups, dichotomized using the median risk score computed on the training dataset. We computed the stage-adjusted hazard ratios and Wald test *P*-values by comparing the survival data of the low-risk versus high-risk patients identified across the testing dataset. Only the gene signature and median risk score were carried over from the analysis of the training dataset, making this testing dataset fully independent.

For the second lung adenocarcinoma benchmark, we used the Beer cohort [40] as a training dataset and all four cohorts from the Director's Challenge as a testing dataset [41], pre-processed and evaluated as described above.

Prediction of extra-prostatic extension in prostate tumors

As another prediction task, we used prostate tumor expression profiles to predict extra-prostatic extension (EPE) of the prostate tumors, which is a strong predictor for recurrence [45]. Clinically annotated gene expression profiles for 69 prostate tumor and 18 healthy prostate samples were downloaded [43], normalized with RMA, and pre-processed and purified by ISOpure as described in the previous section. EPE is a binary outcome indicating whether or not extension has occurred, therefore, prediction of EPE in prostate tumors is a classification problem. We trained elastic net-regularized logistic regression classifiers using either the original expression

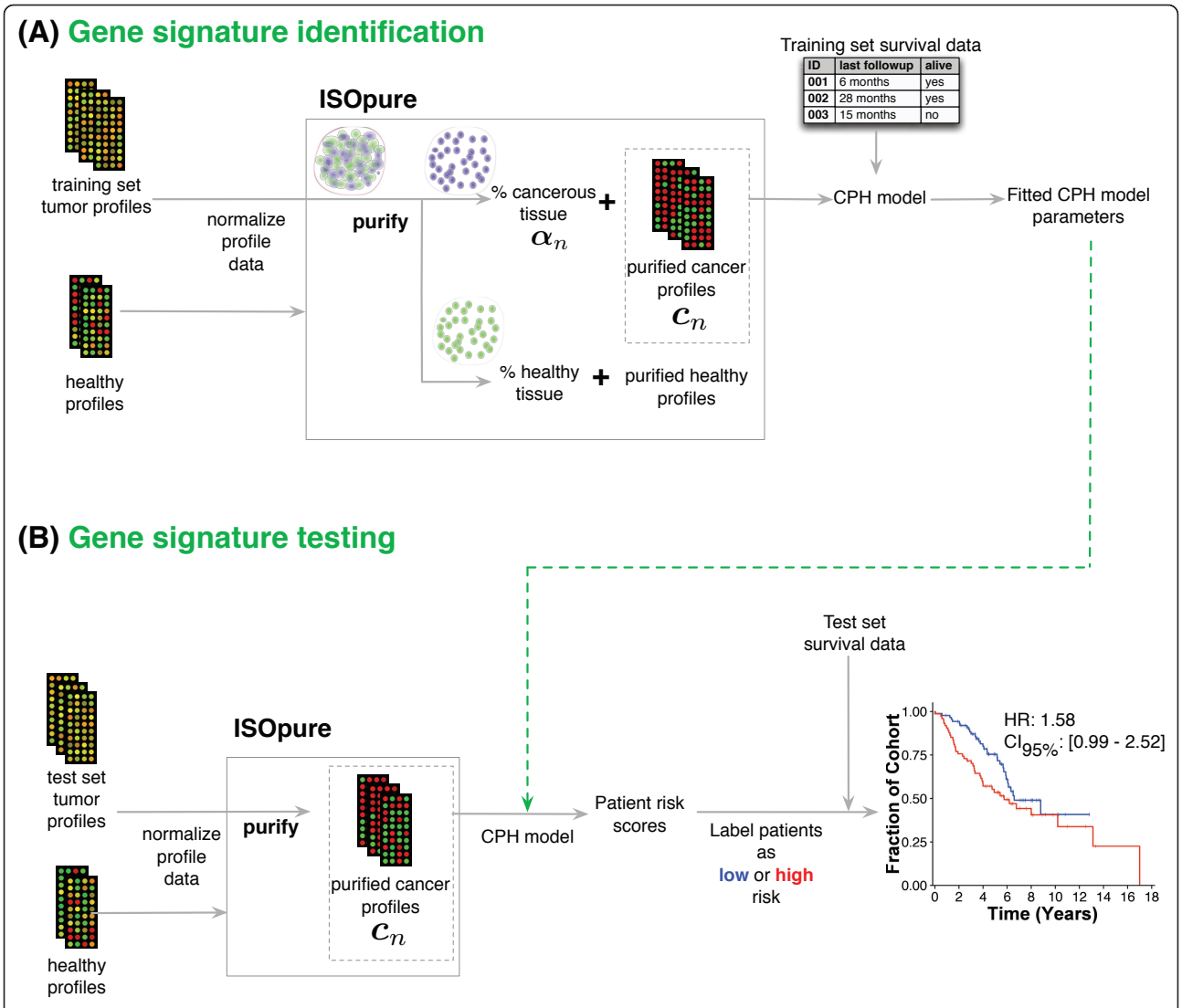

Figure 3 Pipeline for tumor purification and subsequent identification and testing of gene expression-based prognostic models. **(A)** Gene signature identification stage. First, profiles from tumor samples and healthy tissue are co-normalized together using the robust multi-array average (RMA) method, then input into ISOpure to estimate the purified cancer profiles $\boldsymbol{c}_n$. The purified cancer profiles are used as covariates to train an elastic net-regularized Cox proportional hazards (CPH) model (the gene signature) to predict survival data associated with each tumor sample. The trained parameters of the CPH model are used later in model testing. **(B)** Gene signature testing stage. First, new (test-set) tumor profiles are co-normalized with healthy tissue profiles and purified using ISOpure. Each purified cancer profile is then used to compute a risk score for the corresponding patient, using the CPH model parameters learned in the training stage. Patients in a test cohort are then divided into low-risk and high-risk groups based on their risk score, and the hazard ratio is calculated to evaluate the low- and high-risk classifications.

profiles, the cancer expression profiles estimated by ISOpure, the estimates of sample composition made by matrix factorization, or the cancer expression profiles estimated by the Clarke method. We used glmnetCV (version 1.3) [44] on R (version 2.11.0) to train each model and to measure the 10-fold cross-validation accuracy. To set the two regularization parameters we conducted a grid search, evaluating each combination of parameters using 10-fold cross-validation on the training set. For each of the 100 λ values selected by glmnet by default, we tried all values of α between 0.1 and 0.9, inclusive, with step size of 0.1. We then trained a model using the entire training set with the regularization parameters that led to the highest training accuracy during the grid search. To reduce the effect of fold selection, and allow pairwise comparisons of prediction accuracy between each purification method, each method was trained with identical fold divisions. The entire procedure was repeated 10 times, and the mean accuracy on the held-out folds is reported.

Measuring prognostic performance as a function of training cohort size

To measure gene signature performance as a function of training cohort size, we repeated the gene signature testing procedure described above using the first lung adenocarcinoma benchmark (that consists of separate training and testing datasets from the Director's Challenge cohorts), except that we sub-sampled (without replacement) tumor samples from the training dataset. For each training cohort size tested, we constructed prognostic models on 1,000 random subsets of the full training dataset, and evaluated model performance on the 186 patients in the test-set cohort. We tested training cohort sizes of 50 to 250 patients, in increments of 10 patients. We used linear interpolation between tested cohort sizes to estimate model performance on other training cohort sizes.

Results

The ISOpure algorithm is outlined in Figure 1. To evaluate ISOpure, we compared the predictive performance of prognostic models (gene signatures) generated from the original unpurified microarray expression profiles with models generated from the cancer profiles estimated by ISOpure and the Clarke methods, and the mixture proportion profiles from matrix factorization. Figure 3 and Methods outline our procedure for tumor purification, identification of a prognostic gene signature on a training cohort, and testing the gene signature on an independent patient cohort. We selected two tumor types for this evaluation (prostate and non-small cell lung adenocarcinomas) based on the availability of large cohorts [41], and because these tumor types do not yet have established sub-types. We selected prognostic prediction as the clinical task, because in both diseases, treatment escalation/de-escalation is of immediate clinical relevance. The majority of intermediate-risk prostate cancer patients are over-treated, and current therapies such as prostate removal result in serious morbidities. It is predicted that up to a quarter of patients with non-small cell lung cancer would derive benefit from treatment escalation but do not receive it, whereas a similar number of patients classified as stage II are thought to be over-treated [9,46].

Computational purification improves prognostic gene signatures for lung and prostate cancer

We compared the predictive performance of the unpurified and the ISOpure cancer profiles on the Director's Challenge [41] benchmark of 440 lung adenocarcinomas collected in four cohorts from four different institutions (Figure 4). This benchmark used two of the cohorts (N = 254 patients) as a training set, and two other cohorts (N = 186 patients) as a held-out, independent test set. We identified separate gene signatures (CPH models) on the unpurified profile and the ISOpure cancer profile training sets, and used them to classify test-set patients into low-risk and high-risk groups. Model performance was measured by the hazard ratio (HR; that is, the relative hazard of death for samples classified in the high-risk group), with a higher value indicating better performance. The ISOpure-based signature (HR = 2.92, $P = 3.47 \times 10^{-5}$, Wald test) was significantly better at predicting patient outcome (P = 0.001, likelihood ratio test) than the unpurified profile-based signature (HR = 2.01, P = 0.006, Wald test). Note that the unpurified profile-based signature is an extremely strong baseline for comparison: none of the eight groups in the Director's Challenge generated a gene signature that was significantly better than random. Purifying tumor profiles using representatives of existing expression deconvolution methods (the Clarke method [33] and a matrix factorization-based method [14]) degraded performance compared with the unpurified profiles (HR_{clarke} = 1.83, HR_{mf} = 1.09) (see Additional File 8: Figure S2).

To introduce technical variability into the tumor profiles, as may arise in clinical conditions, we repeated our evaluation procedure using a training set [40] collected on a different platform and in a different study from that of the test-set data [41] (Figure 5). This also allowed us to use the full Director's Challenge cohorts (one of the largest lung adenocarcinoma datasets available) as a test set. Again, we found the test-set performance of the ISOpure-based signature of 110 genes (ISOpure-sig; HR = 1.87, $P = 4.7 \times 10^{-6}$, Wald test) was significantly better ($P = 2.77 \times 10^{-4}$, likelihood ratio test) than the 82-gene signature based on the unpurified profiles (unpurified-sig; HR = 1.48, P = 0.004, Wald test). These two signatures (see Additional File 9: Table S1; see Additional File 10: Table S2) have a significant overlap of 48 genes ($P = 2.0 \times 10^{-61}$, Fisher's exact test), all of which are in agreement about whether their expression increases or decreases the hazard for death. Note that the larger number of genes in ISOpure-sig is not an indication of bias in favor of that method; the size of each signature is selected automatically by a cross-validation procedure that attempts to maximize training-set performance (see Methods). Expanding the size of the unpurified-sig gene signature to 110 genes decreased its test-set performance (HR = 1.41, P = 0.011, Wald test). The improved performance of ISOpure is due to the novel regularization used in the second step of the algorithm: the ISOpure-evenprior model, which replaces the Dirichlet prior for $\boldsymbol{c}_n$ used in the second step of ISOpure with a different, commonly used Dirichlet prior, performs comparably to unpurified-sig (HR = 1.51, P = 0.004, Wald test).

Of the 440 test-set samples, 70 were classified differently by ISOpure-sig and unpurified-sig. These 70 samples are significantly enriched for stage IB tumors (P = 0.011,

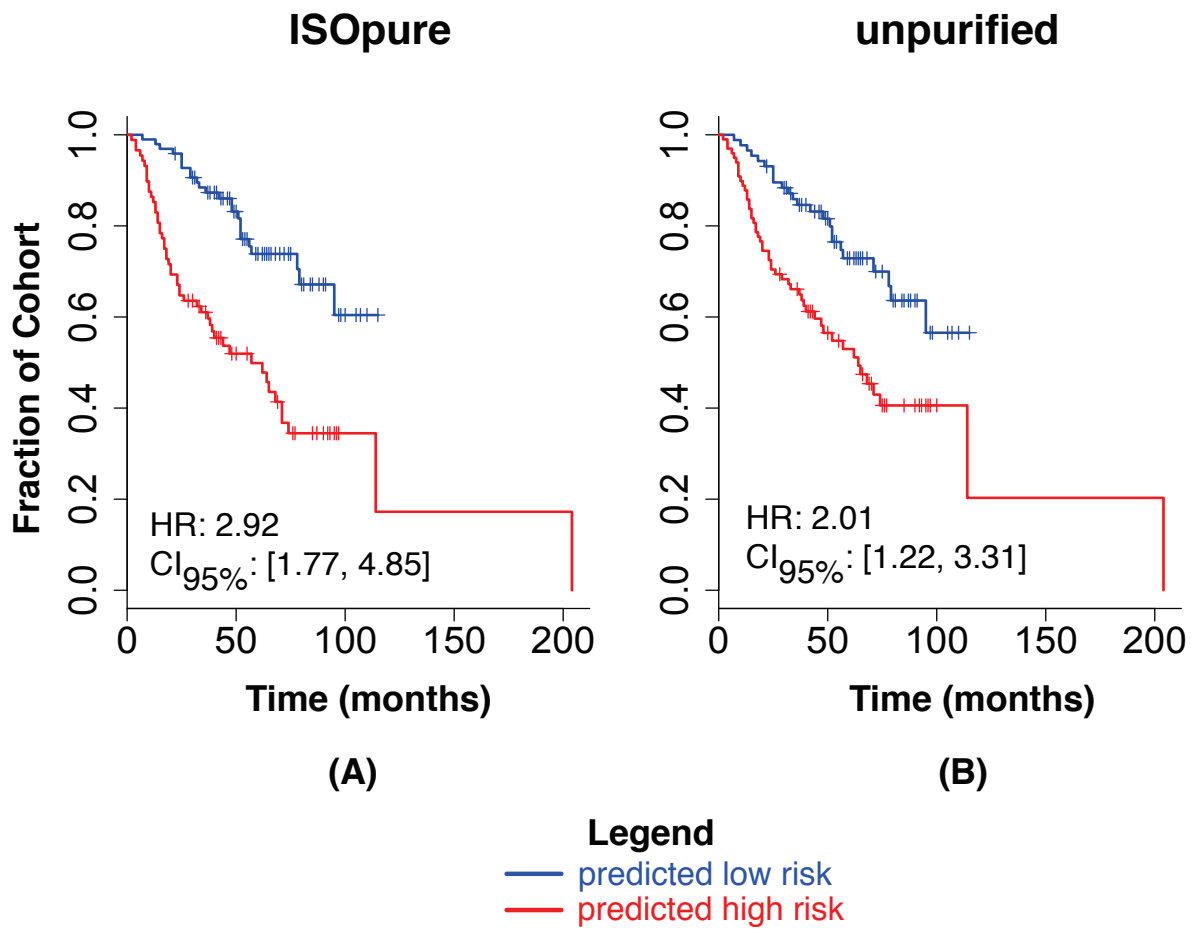

Figure 4 Test-set performance of CPH models on the MSKCC and DFCI cohorts of the Director's Challenge. Each plot shows Kaplan-Meier survival curves for the predicted low-risk group (blue) and the predicted high-risk group (red). Plots show the empirical survival probability *S(t)* as a function of time. The hazard ratio (HR) of the high-risk versus the low-risk groups is given, along with its 95% confidence interval. Test-set performance is reported for both the **(A)** ISOpure cancer profiles and **(B)** original, unpurified profiles.

Bonferroni-corrected hypergeometric test) (see Additional File 11: Figure S3). When the testing cohort was restricted to stage IB tumors, ISOpure-sig provided an even greater benefit over unpurified profiles ($HR_{ISOpure\text{-}sig}$ = 2.33, $HR_{unpurified\text{-}sig}$ = 1.58, P = 0.003, likelihood ratio test) (Figure 5). In non-small cell lung adenocarcinomas, the predicted prognosis of early-stage tumors influences the decision of whether to perform adjuvant therapy [46], so these improvements in prognostic prediction for early-stage tumors are relevant to improving patient care. Note also that stage I tumors in this cohort had significantly lower predicted cancer content than those of later stages (see Additional File 12: Figure S4), which may explain the larger difference in performance. We further verified that our ISOpure-sig model was a significant predictor of outcome across all patients with stage I cancer (HR = 1.73, P = 0.005, Wald test) (see Additional File 13: Figure S5).

Next, we evaluated ISOpure on prostate tumor data. In prostate cancer, the presence of EPE is a strong predictor for recurrence [45], and also indicates the need for post-operative radiotherapy to maximize survival [47]. Current guidelines for the evaluation of EPE are subjective [48] and can only be applied post-operatively. Accurate, objective assessment of EPE based on biopsies would contribute to optimal patient treatment by prioritizing patients for prostate removal.

Gene expression data from the Wallace study [43] for 69 tumor and 18 normal prostate samples were purified using ISOpure, then used to predict EPE. Because of the small number of samples in the dataset, and the lack of a separate test dataset, we used 10 rounds of 10-fold cross-validation to assess relative performance (Table 2). Classifiers trained on the ISOpure cancer profiles were significantly more accurate than classifiers trained using the original unpurified profiles, the Clarke cancer profiles, or the matrix factorization mixture estimates (all pairwise P-values < 0.005, Wilcoxon signed rank test). Additionally, prediction accuracies were not significantly

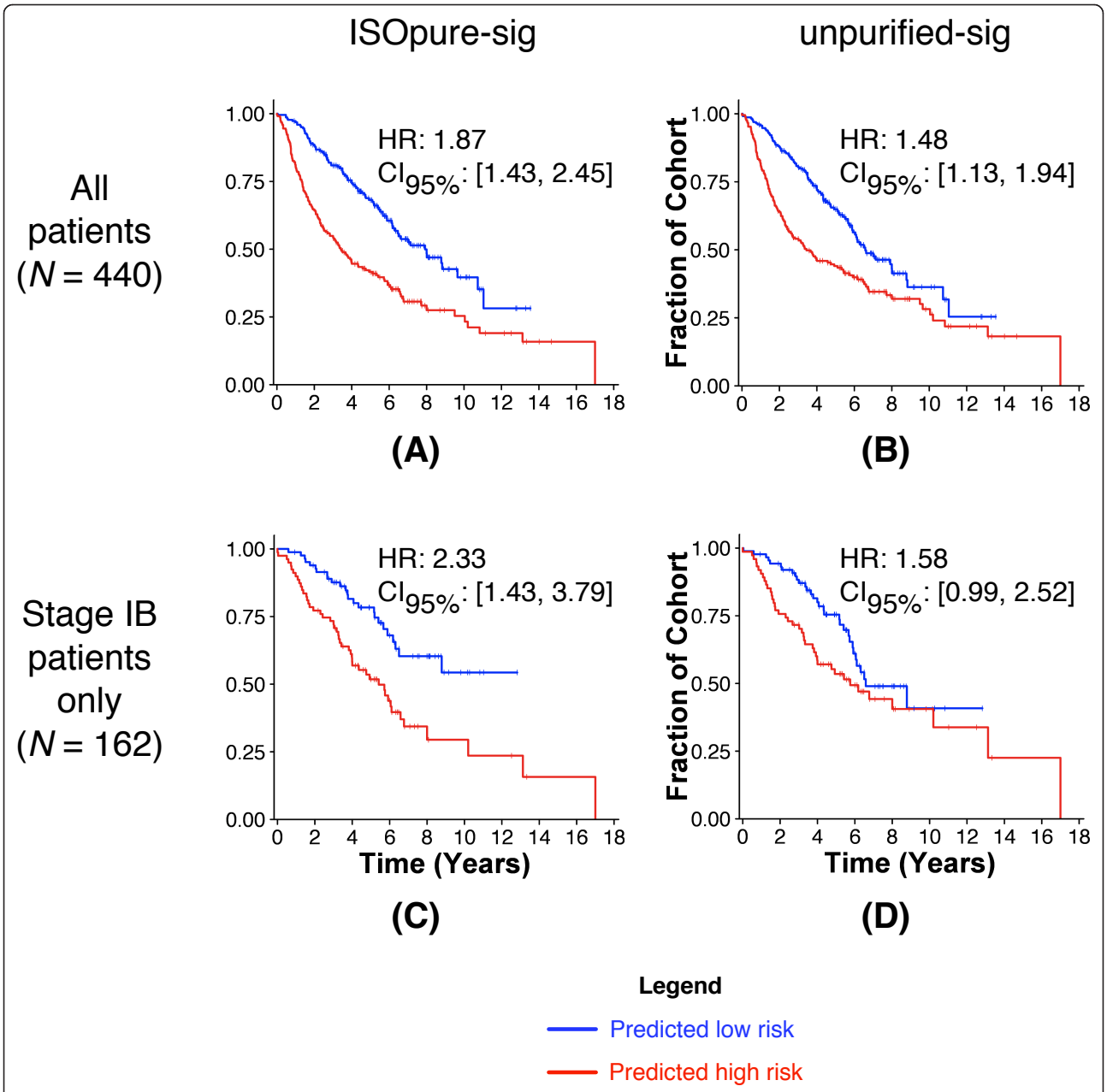

Figure 5 Test-set performance of CPH models on all four cohorts of the Director's Challenge. Plots show Kaplan-Meier survival curves as in Figure 4, using CPH models trained on the Beer [40] cohort. Test-set performance of **(A)** ISOpure-sig classification of the ISOpure cancer profiles from the Director's Challenge cohorts and of (**B**) unpurified-sig classification of the unpurified profiles from the Director's Challenge cohorts. **(C)** Same as (A), but the performance is measured only for patients with stage IB cancer. **(D)** Same as (B), but the performance is measured only for patients with stage IB cancer.

different between the three non-ISOpure methods (all pairwise *P*-values > 0.4).

ISOpure tumor purity predictions correlate with pathologist estimates

Several tumor datasets provide pathologist estimates of tumor cellularity. We used these estimates as a benchmark for the ISOpure estimate of tumor purity. Note that under accurate purification, we expect tumor purity and tumor cellularity to be correlated but not equal, as tumor purity is an estimate of the proportion of mRNA in the sample contributed by cancerous cells, whereas tumor cellularity is based on area and cell counts. Both cell size and the amount of mRNA per cell can vary considerably between cancer and normal cells [49]. Furthermore, pathologists typically assess a slide of the tumor adjacent or proximal

Table 2 Accuracy[a] of elastic net-regularized models for the prediction of extra-prostatic extension (EPE).

Classifier input	Average accuracy, %
Unpurified expression profiles	61.76 ± 1.64
ISOpure cancer expression profiles	69.12 ± 0.90
Matrix factorization estimates	62.94 ± 0.57
Clarke cancer expression profiles	62.50 ± 1.06

[a]The average accuracy and standard error of the mean over ten 10-fold cross-validation runs are reported for logistic regression classifiers trained using either the original unpurified profiles, ISOpure cancer profiles, matrix factorization mixing proportions, or Clarke cancer profiles.

to the sample processed for molecular analysis, whereas ISOpure assesses the purity of the sample processed for molecular analysis [21,22].

On a dataset of 20 lung adenocarcinomas and three blinded control samples from Bhattacharjee and colleagues [22] (see Methods), the ISOpure estimates were well correlated (Spearman's ρ = 0.51; P = 0.013) with the average of two pathologists (Figure 6A; see Additional File 14: Table S3), although the correlation was lower when the three control samples (that are correctly assigned zero cellularity by ISOpure) were removed (Spearman's ρ = 0.26; P = 0.28). Computational purification reduced the inter-tumor variance of all 8,193 gene expression levels in the Bhattacharjee dataset, as expected (Figures 6B, C; see Additional File 15: Figure S6) ($P < 2.2\text{x}10^{-16}$, Wilcoxon signed rank test).

We also found high correlation between ISOpure and a pathologist on a dataset of 109 prostate samples (Figure 6D, Spearman's ρ = 0.75; $P = 1.2 \times 10^{-20}$; see Additional File 16: Table S4), although there was a number of samples for which either ISOpure or the pathologist estimated zero cancerous content but the

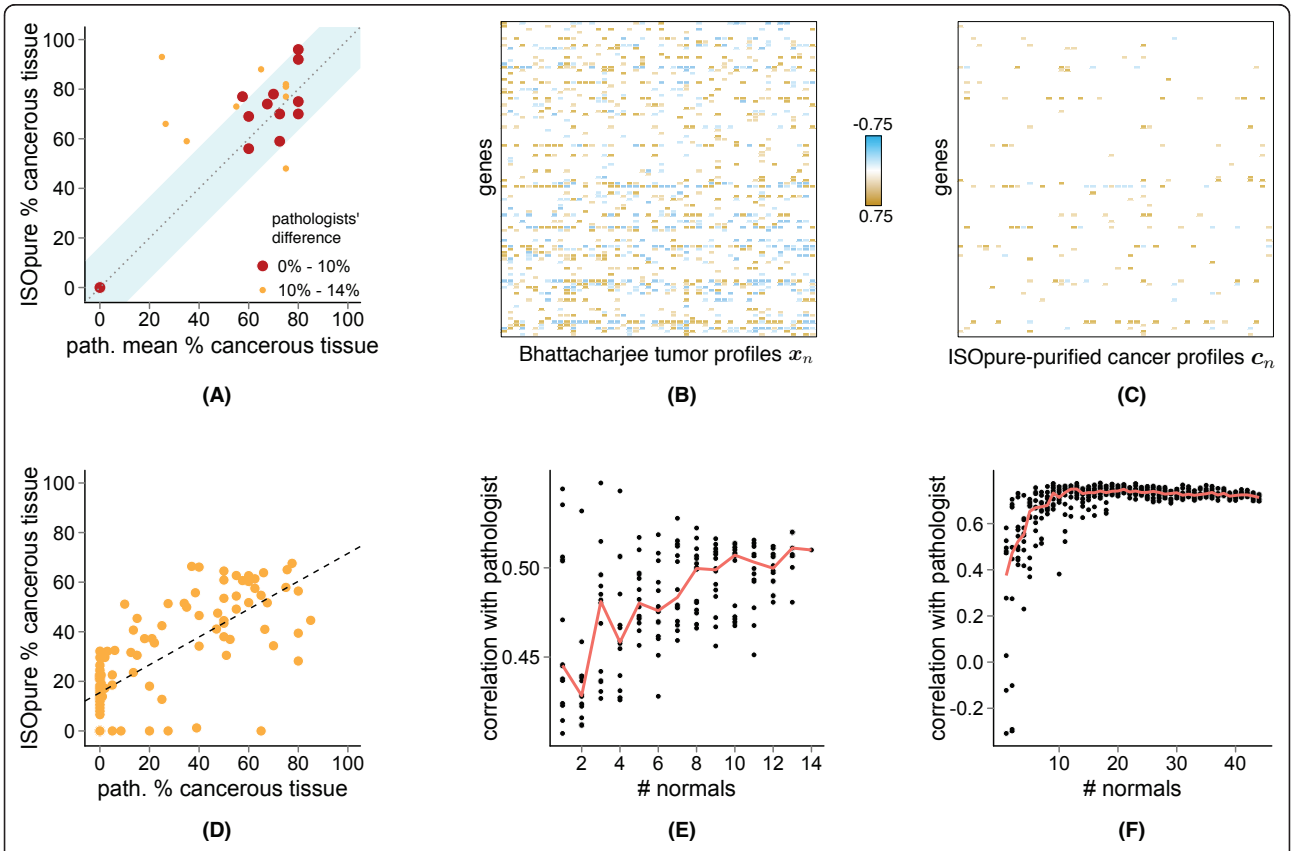

Figure 6 Comparison of ISOpure-predicted and pathologist reported percentage cancerous tissue. **(A)** Scatter plot of ISOpure predictions against the average pathologist estimates on a subset of 20 lung tumors and three blinded healthy lung tissues from the Bhattacharjee dataset. The size and color of each point indicate the difference between the pathologists' estimates. The blue region indicates where the ISOpure predictions were within 13.7% of the average pathologist estimate. **(B)** Median-centered expression levels of a random selection of 100 genes in 50 patients of the Bhattacharjee dataset, before ISOpure purification. **(C)** Same genes and patients as in (B), but expression levels were from the ISOpure cancer profiles. **(D)** Scatter plot of ISOpure predictions against a single pathologist on the Wang dataset of 109 prostate tumor samples. The black dashed line indicates the linear regression model that minimizes the sum of squared errors. **(E)** Correlation of ISOpure estimates and the average of the two pathologists' estimates on the same 23 samples as in (A), depicted as a function of the number of normal samples made available to ISOpure. Each point represents a random selection of normal samples of the given size. **(F)** Same as in (E), but correlation was measured for the 109 samples in the Wang dataset.

other estimated non-negligible (> 20%) cancerous content. However, predictions made by an independent computational method seemed to be consistent with ISOpure for the tumor samples assigned zero cellularity by ISOpure (see Figure S1B in Wang *et al.* [21]). Furthermore, all of the samples assigned zero cellularity by the pathologist were from surgically removed prostates that were initially diagnosed as cancerous; these assessments of zero cellularity may result from field effects [50] or from the tumor-adjacent slides having negligible cancer content in low-cellularity tumors. As we observed in the lung adenocarcinoma dataset, ISOpure reduced the inter-patient variance in expression for all genes ($P < 2.2 \times 10^{-16}$, Wilcoxon signed rank test; see Additional File 17: Figure S7).

For both lung and prostate cancer, having a larger set of normal profiles available for purification improved correlation between ISOpure and pathologist cellularity estimates (Figures 6E, F). Correlation for both tumor types saturated at approximately 10 normal profiles. Even with only one normal profile, the ISOpure median correlation was still higher than that of the Clarke method on both the lung (Spearman $\rho_{\text{ISOpure}} = 0.45$, $\rho_{\text{Clarke}} = 0.35$) and the prostate (Spearman $\rho_{\text{ISOpure}} = 0.38$, $\rho_{\text{Clarke}} = 0.19$) datasets. Note that matched normal samples were not available for every tumor, as assumed by the Clarke method, so instead we matched each tumor with its most highly correlated normal sample.

Discussion

Computational purification of tumor expression profiles by ISOpure improves the accuracy of subsequent prognostic models for lung and prostate cancer by reducing inter-sample variation in the amount and type of gene expression signal in the tumor profile that is due to normal tissue contamination. Purifying tumor profiles using other algorithms did not yield classification performance significantly better than the original unpurified profiles (see Additional File 8: Figure S2). Appropriate regularization, embodied by ISOpure priors, was a key factor in this improvement in accuracy; we observed decreased prognostic accuracy with both an unregularized purification method (Clarke) and a modified version of ISOpure with a standard (but inappropriate) prior (see Additional File 18: Figure S8). This modified version of ISOpure used the same tumor purity estimates as ISOpure, and therefore this result indicates that accurate prediction of tumor purity alone is not sufficient for improving prognostic accuracy. Note that although computational purification is related to the well-studied area of expression deconvolution, standard expression deconvolution algorithms do not support profile-specific purification because the only profile-specific information they generate is the tumor purity (more specifically, the proportions of a handful of inferred or provided cell-type specific profiles) [14,20,21,26-32], and these proportions are not strong prognostic indicators in either prostate or lung cancer (Table 2; see Additional File 8: Figure S2). We designed ISOpure as a mixture model that models each tumor sample as a mixture of sample-specific normal and cancer profiles (with some constraints implied by the representative cancer profile and the healthy profiles). Thus, despite substantial efforts in this area, ISOpure is the first validated expression deconvolution algorithm that satisfies a set of reasonable requirements for clinical use of computational purification (Table 3).

During preparation of the final revision of our manuscript, we were introduced to the disease-specific genomic analysis (DSGA) algorithm [51], which could, in theory, be adapted to computational purification. DSGA, like ISOpure, models $\boldsymbol{h}_n$ as a linear combination of a set of provided normal profiles but, unlike ISOpure, does not regularize its estimate of $\boldsymbol{c}_n$. Thus, we suspect that its performance on our benchmarks would be similar to that of ISOpure-evenprior.

Note that the ISOpure regularization strategy is a compromise between previous methods that either over-regularize (and assume all tumor samples are composed of the same small number of cancer and normal cells) or do not regularize (and therefore attempt to solve an ill-posed statistical problem). However, regularization entails making specific assumptions about the nature of the purified cancer expression profiles, and purification quality would probably be reduced if the regularization assumptions were violated.

ISOpure makes two key assumptions. First, it assumes that the set of provided normal profiles contains representative samples of the profiles of the contaminating normal tissue. This assumption could be violated if the set of provided normal profiles does not contain sufficient samples of the type of tissue contaminating the tumors, or if uncorrected batch effects lead to systematic differences in expression between the set of normal profiles and the set of tumor profiles that are not due to cancer. The normal profiles used in the current study were selected to minimize batch effects and we have not evaluated batch-correction procedures with ISOpure. Note, however, that we did not require the profiled normal tissue samples to be from the same patient, and most of the tumor samples we used did not have matching normal profiles available. For prostate and lung, purification using between 10 and 30 normal profiles seemed sufficient; correlation with pathologist estimates of tumor cellularity stopped improving after 10 samples for both tumor types, although the predictive performance of lung cancer prognostic models continued to improve until 30 profiles of normal lung tissue were used (see Additional File 19: Figure S9). The accuracy of

Table 3 Evaluation of the suitability of ISOpure and other expression deconvolution methods for clinical use.[a,b]

Method	Estimates individual cancer profiles	Uses unmatched normal tissues	Requires minimal additional data	Tested on clinical data
ISOpure	Yes	Yes	Yes	Yes
ISOLATE [14]	No	NA	Yes	No
Erkkila [28]	No	NA	Yes	No
Lahdeskmaki [27]	No	NA	Yes	No
Venet [26]	No	NA	Yes	No
Tolliver [31]	No	NA	Yes	No
Ghosh [29]	No	Yes	No	No
Shen-Orr [30]	No	NA	No	No
Bar-Joseph [32]	No	NA	No	No
Wang [21]	No	NA	No	No
Stuart [20]	No	NA	No	No
Gosink [34]	Yes	No	Yes	No
Clarke [33]	Yes	No	Yes	No

[a]Methods require minimal additional data if they do not require mixing proportions of normal and cancer cells for each input tumor sample for deconvolution.
[b]NA, not applicable; No, negative assessment; Yes, positive assessment.

the prediction of EPE increased steadily with the number of normal samples available for purification, even when we reached 18 normal samples, the maximum number available (see Additional File 20: Figure S10). These results suggest that collection of more normal samples may have further improved prediction of EPE, although only two normal samples were required to improve EPE prediction significantly above the baseline. In general, our observations suggest that just one or a small number of normal samples will inadequately represent the biological variability in normal gene expression, and collecting as many as 30 normal samples may be necessary to adequately capture normal variation.

ISOpure also assumes that the tumor-specific cancer profiles are similar to the representative cancer profile, $\boldsymbol{m}$, estimated in the first step. This assumption is appropriate for sets of tumor profiles without strong expression sub-types, such as the prostate and lung datasets we used, but may be violated and lead to decreased performance for cancers with distinct expression sub-types, such as breast cancer. For these cancers, we recommend grouping the tumor profiles by subtype and purifying each group separately. Note that it is possible to extend ISOpure to consider multiple sub-types in the cohort by allowing multiple clusters of cancer profiles. However, doing so would require the estimation of a different '$\boldsymbol{m}$' vector for each cluster, so in order to avoid over-fitting, the number of inferred sub-types must be much smaller than the number of tumor profiles in the dataset.

The performance of ISOpure is relatively robust against small perturbations in its inputs. With respect to correlation of ISOpure tumor purity estimates with pathologists, the median correlation of ISOpure estimates decreases by 0.01 when using 12 versus 13 normal samples for purifying the lung dataset, whereas the median correlation actually increases by 0.006 when using 43 versus 44 normal samples for purifying the prostate dataset. In terms of prognostic prediction performance, the median accuracy of the ISOpure cancer profiles decreases by 0.015 when using 17 instead of 18 samples for the prostate tumors.

ISOpure-sig is the first validated prognostic signature for the well-studied Director's Challenge benchmark, and it is also prognostic for patients with stage I cancer alone, a group for which good prognostic models are urgently needed for clinical application. The excellent performance of ISOpure for this sub-group may result from the significantly lower cancer content of stage I tumors compared with later stage tumors. This suggests that ISOpure would have similar performance gains on other samples with low cancer content. In addition to improving accuracy for a given patient cohort size, ISOpure can be used to increase the effective size of a patient cohort by reducing inter-patient variability due to contamination. This reduces the cost of cancer biomarker studies and, crucially, enables gene signature identification and application for tumor types for which fewer tumor samples are available (see Additional File 21: Figure S11). We have shown the utility of ISOpure for 834 samples derived from five datasets of two entirely different tumor types. Nevertheless, it is possible that unique features of other diseases or data types will change performance characteristics in different situations, and therefore additional and on-going validation in emerging large datasets [52,53] will be essential.

Our analysis demonstrated approximately 10% improvement in prediction of EPE when using ISOpure cancer

profiles compared with the unpurified profiles. EPE prediction is a challenging problem, and we note that without purification, the predictive model performance was only as accurate as simply picking the majority class, so the 10% improvement was actually an increase from zero benefit of considering unpurified expression profiling data. Prostate cancer is the most common malignancy in men, and treatment is often determined entirely by risk groups assigned using pre-treatment prostate-specific antigen levels, biopsy-based Gleason scores, and T category. As a result, improved prediction of EPE from biopsies (that is, improved estimates of T category) could provide benefit when combined with these other risk predictors.

Two requirements must be met in order for ISOpure to be applicable in the clinic. First, collection of gene expression profiles for each patient's tumor sample must become part of the medical diagnosis pipeline. To that end, gene sequence or expression profiling is already being used to inform treatment decisions for multiple cancer types, including breast, gastric, lung, and colorectal cancer [54]. Second, ISOpure relies on expression profiles of normal tissue samples to remove contamination from tumor samples. In the Gene Expression Omnibus (GEO), the number of tumor datasets with associated normal samples is a small subset of the total number of tumor datasets. Collection of a database of normal samples from multiple tissue sites will be needed to use ISOpure; however, as shown by our deconvolution of the Director's Challenge cohort of 443 samples, deconvolution can even use existing datasets of normal samples, if the collection protocol and platform are sufficiently similar.

Although our analysis here focuses on microarray expression profiles, we speculate that ISOpure could be applied directly to abundance estimates from sequencing (RNA-seq) data. Many methods for analyzing RNA-seq transform the reads into a vector of fixed length, whose elements represent abundance estimates, such as RPKM (reads per kilobase per million mapped reads) for genes [55], transcripts (including splice isoforms) [56], or individual exons and exon-exon junctions [57]. These abundance estimates could be directly input into ISOpure as if they are microarray expression profiles, possibly after rescaling them to increase the precision of the discretization of these profiles into count vectors. ISOpure is, however, currently unable to take advantage of data on somatic genetic variants that may help to distinguish reads from normal and tumor RNA. The value of this genetic data may increase with increasing read lengths because the chance that an individual read will cover a polymorphic region will also increase. Future extensions of ISOpure could include these data either as priors on estimates of tumor cellularity or directly as part of the generative model.

Although in this work we focused on addressing inter-tumor heterogeneity due to normal tissue contamination, another source of tumor expression variability is intra-tumor heterogeneity. We expect that when individual tumors contain more than one cancer cell state, the estimated cancer profile of ISOpure would be a weighted average of the different cancer cells contributing to the provided expression profile. However, typical dataset sizes make estimation of even a single cancer profile per tumor sample extremely challenging, and was the focus of ISOpure development. We expect that until multiple samples from each individual tumor can be obtained in a widespread manner, it will not be feasible to address the problem of estimating multiple cancer profiles for each tumor sample because of the small sample sizes. Nonetheless, even with multiple samples per tumor, normal contamination will still be a problem, necessitating the use of tools such as ISOpure. We further suggest that removing the influence of normal contamination may make it easier to distinguish expression patterns unique to sub-clonal populations within the purified profile.

Sample purification increases the number of patients that can benefit from prognostic models by rescuing samples that otherwise would be discarded because of low cellularity. We and others [20-23,43] have found that tumor samples selected for gene expression analysis vary widely in their cancerous tissue content (Figure 7; see Additional File 14: Table S3; see Additional File 16: Table S4), so a large number of patients stand to gain from improved sample purification. By using ISOpure for purification, prognostic predictions can be made immediately after molecular profiling, at minimal marginal cost. ISOpure can also play an important supplemental role to pathological evaluation of tumors by providing an independent assessment of cancerous RNA content.

Conclusions

We report a computational purification tool, ISOpure, which mitigates the effects on gene expression profiles of normal tissue contamination in tumor samples and leads to significant improvement in the prediction of patient prognosis and other clinical variables in lung and prostate cancer. The purification, gene signature identification, and testing procedures presented here are fully automated and unbiased, and require only tumor and healthy tissue samples, and associated clinical data (for example, survival or progression indicators). Our procedure can therefore complement any gene signature identification method for any solid tumor, possibly also including those focusing on RNA-seq, protein abundance, or DNA copy-number variation. Although we have chosen here to build upon the architecture of the

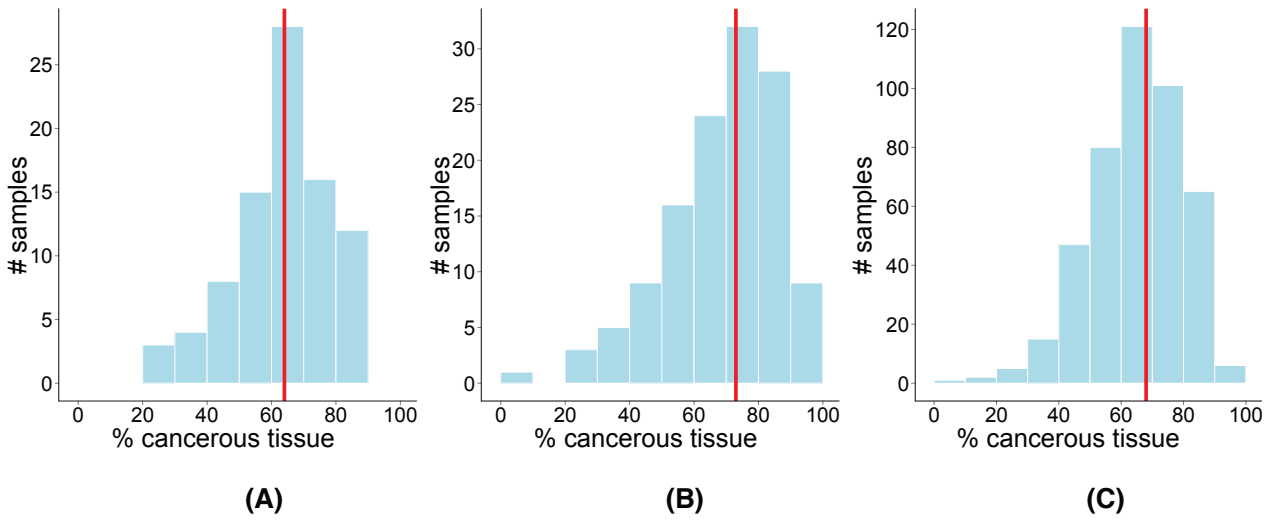

Figure 7 ISOpure percentage cancerous tissue estimates for 656 lung adenocarcinomas from three datasets. The median percentage cancerous tissue is shown as a red solid line. **(A)** The Beer dataset (*N* = 86); **(B)** the Bhattacharjee dataset (*N* = 127); (**C**) the four cohorts from the Director's Challenge dataset (*N* = 443).

existing deconvolution algorithm ISOLATE using one particular regularization strategy, other regularization strategies and deconvolution methods may be extended to provide purified, per-tumor cancer expression profiles as well. Although ISOpure has demonstrated success for the analysis of lung adenocarcinoma and prostate tumor samples, future development may be needed to incorporate the possibility of multiple sub-types in a single patient cohort. Nonetheless, we have shown that computational purification methods can improve downstream analyses of tumor expression profiles. We therefore conclude that more exploration of intermediate-strength regularization strategies such as ISOpure may yield significant improvement in downstream analyses of tumor samples and other situations in which biological samples are composed of mixtures of cell types.

Additional material

Additional File 1: ISOpure MATLAB code (ZIP format).

Additional File 2: ISOpure cancer profiles from the Beer and Bhattacharjee cohorts, and part of the Shedden cohorts (ZIP format).

Additional File 3: ISOpure cancer profiles from part of the Shedden cohorts (ZIP format).

Additional File 4: ISOpure cancer profiles from part of the Shedden cohorts (ZIP format).

Additional File 5: ISOpure cancer profiles from the Wallace cohort (ZIP format).

Additional File 6: Implementation of Clarke's method for estimating tumor purity (R code file).

Additional File 7: Figure S1: Comparison of percentage cancerous tissue made by each pathologist on the Bhattacharjee dataset (PDF file). The dotted line indicates the $y = x$ axis, and the blue region indicates where the difference between the estimates of the two pathologists is less than 13.7% (one standard deviation of their overall differences).

Additional File 8: Figure S2: Test-set performance of CPH models on the MSKCC and DFCI cohorts of the Director's Challenge (PDF file). We followed the pipeline presented in Figure 3 to train and test gene signatures. We used the Director's Challenge training and testing cohorts as defined in the original study. Illustrated are the test-set performances of CPH models based on **(A)** ISOpure cancer profiles, **(B)** original, unpurified tumor profiles, **(C)** Clarke cancer profiles, and **(D)** matrix factorization mixing proportions (the 50 mixing weights of the cancer and normal profiles estimated by ISOpure Step 1). Performance is adjusted for pathological stage.

Additional File 9: Table S1: Entrez ID and weight of each gene in the 82-gene signature derived from the original, unpurified lung tumor profiles of the Beer cohort (unpurified-sig) (XLS file).

Additional File 10: Table S2: Entrez ID and weight of each gene in the 110-gene signature derived from the ISOpure lung cancer profiles of the Beer cohort (ISOpure-sig) (XLS file).

Additional File 11: Figure S3: Stage-wise stratification of the patients who were differentially and similarly classified by ISOpure-sig and unpurified-sig (PDF file). **(A)** Plot shows the stratification of the 70-patient sub-group classified differently ('diff') by ISOpure-sig and unpurified-sig, and the entire group ('all'). The number of patients in each category is shown above each bar. **(B)** Same as (A), but showing those 370 patients similarly classified between the two signatures ('same').

Additional File 12: Figure S4: Distributions of percentage cancerous tissue for patients with stage I cancer versus all other stages, computed over all three lung adenocarcinoma datasets (PDF file). *N* indicates the number of samples plotted in each box. Six samples were excluded because of missing stage information.

Additional File 13: Figure S5: Test-set performance of CPH models on the 277 stage I patients from the Director's Challenge (PDF file). In these prediction experiments, the prognostic models are trained on the Beer cohort, and tested on the stage I patients from the Director's Challenge cohorts. Performance of the prognostic models is based on **(A)** the original unpurified profiles and **(B)** the ISOpure cancer profiles.

Additional File 14: Table S3: Estimates of percentage cancerous tissue made by ISOpure on all three lung adenocarcinoma (Bhattacharjee, Beer, Shedden) datasets (XLS file).

Additional File 15: Figure S6: Inter-patient variance of expression levels for 8,193 genes in the Bhattacharjee dataset, before and after ISOpure purification (PDF file). The red dashed line is the $y = x$ line (no change in variance).

Additional File 16: Table S4: Estimates of percentage cancerous tissue made by ISOpure on the Wang prostate dataset (XLS file).

Additional File 17: Figure S7: Inter-patient variance of expression levels for 18,185 genes in the Wang dataset, before and after ISOpure purification (PDF file). The red dashed line is the $y = x$ line (no change in variance).

Additional File 18: Figure S8: Test-set performance of a CPH model based on either the unpurified profiles, ISOpure cancer profiles, or ISOpure-evenprior cancer profiles (PDF file). ISOpure-evenprior cancer profiles are generated using the same model as ISOpure, except that the Bayesian prior over each individual cancer profile is replaced by a prior whose mean vector is the uniform distribution. **(A)** Test-set performance of a CPH model trained using the Beer cohort and tested on the entire Director's Challenge dataset, when using the original, unpurified tumor profiles. **(B)** Same as (A), but using the ISOpure cancer profiles. (**C**) Same as (A), but using the ISOpure-evenprior cancer profiles. **(D)** Test-set performance of a CPH model trained using the HLM and MI cohorts from the Director's Challenge and tested on the MSKCC and DFCI cohorts from the Director's Challenge, when using the original, unpurified profiles. **(E)** Same as (D), but using the ISOpure cancer profiles. **(F)** Same as (D), but using the ISOpure-evenprior cancer profiles. Performance was adjusted for pathological stage.

Additional File 19: Figure S9: CPH model performance as a function of the number of normal samples for the Director's Challenge dataset (PDF file). We followed the pipeline presented in Figure 3 to train and test a gene signature, using the Beer and Director's Challenge datasets as training and testing cohorts, respectively. The full Beer dataset contains 10 normal samples and the full Director's Challenge dataset contains 49 normal samples from the Landi study. The *x*-axis indicates the maximum number of normal samples available to ISOpure for purifying the tumor samples from the training and testing cohorts. Each box shows the distribution of performance for 49 prognostic signatures, each trained with profiles that were purified using a random subset of normal profiles of the indicated size. Because the training cohort only had 10 normal samples, after $x = 10$ we used all 10 normal samples for purification of the training cohort. The *y*-axis indicates the significance of the improvement in performance over the CPH model trained and tested on the unpurified profiles, as measured by the *P*-value from a likelihood ratio test.

Additional File 20: Figure S10: Improvement in extra-prostatic extension (EPE) predictive performance as a function of the number of normal samples (PDF file). Predictive power improvement was measured as the difference in accuracy between classifiers trained using the original expression profiles and the ISOpure cancer profiles. For each size of the subset of normal profiles tested, 18 random subsets were drawn from the full set of normal profiles.

Additional File 21: Figure S11: CPH model performance as a function of the training cohort size for the Director's Challenge dataset (PDF file). The Director's Challenge cohorts were divided into the same 254-patient training cohort and 186-patient testing cohort used in the original study. Subsets of different sizes of the training cohort were sampled to generate smaller training cohorts, which were then used to identify gene signatures that were evaluated on the full 186-patient testing cohort, as outlined in Figure 3. Results were averaged over 1000 random subsets of each training cohort size and, along with the standard error, are shown for both the ISOpure cancer profiles and the original unpurified profiles. The dotted line indicates the training cohort size required ($N = 212$) for the CPH model based on ISOpure cancer profiles, to achieve the same performance as that achieved by the CPH model based on the original unpurified profiles at a training cohort size of 250 patients. Performance is measured by the hazard ratio (HR), where higher HR is better.

Abbreviations

DSGA: disease-specific genomic analysis; EPE: extra-prostatic extension; HR: hazard ratio; ID: identifier; ISOLATE: Identification of Sites of Origin by Latent Variables; MAP: maximum *a posteriori*; RMA: robust multi-array average; RPKM: reads per kilobase per million mapped reads; SD: standard deviation; CPH: Cox proportional hazards;

Competing interests

The authors have filed for intellectual property protection on the lung cancer prognostic signature ISOpure-sig.

Authors' contributions

GQ, PCB, and QM conceived the project and designed the experiments. GQ, AC, and QM designed the ISOpure model. SH, AGD, and PCB collected and pre-processed the gene expression and clinical indicator data. GQ, SH, AGD, and PCB carried out the experiments. All authors analyzed the data and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This study was funded by a Natural Sciences and Engineering Research Council (NSERC) operating grant and an Early Researcher Award from the Ontario Research Fund to QM, a NSERC PGS Doctoral fellowship to GQ, and a NSERC Julie Payette Scholarship to AGD. This study was conducted with the support of the Ontario Institute for Cancer Research to PCB through funding provided by the Government of Ontario.

Authors' details

[1]Department of Computer Science, University of Toronto, 10 King's College Road, Room 3302, Toronto, ON, Canada, M5S 3G4. [2]Informatics and Biocomputing Platform, Ontario Institute for Cancer Research, 101 College Street, Suite 800, Toronto, ON, Canada, M5G 0A3. [3]Computer Laboratory, University of Cambridge, 15 JJ Thomson Avenue, Cambridge, United Kingdom, CB3 0FD. [4]Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, 10 King's College Road, Room SFB540, Toronto, ON, Canada, M5S 3G4. [5]Division of Engineering Science, University of Toronto, 40 St. George Street, Suite 2110, Toronto, ON, Canada, M5S 2E4. [6]Department of Medical Biophysics, University of Toronto, 610 University Avenue, Room 7-411, Toronto, ON, Canada, M5G 2M9. [7]Department of Molecular Genetics, University of Toronto, 1 King's College Circle, Room 4396, Toronto, ON, Canada, M5S 1A8. [8]The Donnelly Centre, 160 College Street, Room 230, Toronto, ON, Canada, M5S 3E1. [9]Current address: Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA.

Published: 28 March 2013

References

1. Herbst RS, Heymach JV, Lippman SM: **Lung cancer.** *New Engl J Med* 2008, **359**:1367-80.
2. Liotta L, Petricoin E: **Molecular profiling of human cancer.** *Nat Rev Genet* 2000, **1**:48-56.
3. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286**:531-7.
4. Ishibashi Y, Hanyu N, Nakada K, Suzuki Y, Yamamoto T, Yanaga K, Ohkawa K, Hashimoto N, Nakajima T, Saito H, Matsushima M, Urashima M: **Profiling gene expression ratios of paired cancerous and normal tissue predicts relapse of esophageal squamous cell carcinoma.** *Cancer Res* 2003, **63**:5159-64.
5. Korkola JE, DeVries S, Fridlyand J, Hwang ES, Estep AL, Chen YY, Chew KL, Dairkee SH, Jensen RM, Waldman FM: **Differentiation of lobular versus ductal breast carcinomas by expression microarray analysis.** *Cancer Res* 2003, **63**:7167-7175.
6. Boutros PC, Lau SK, Pintilie M, Liu N, Shepherd FA, Der SD, Tsao MS, Penn LZ, Jurisica I: **Prognostic gene signatures for non-small-cell lung cancer.** *P Natl Acad Sci USA* 2009, **106**:2824-2828.
7. Chibon F, Lagarde P, Salas S, Perot G, Brouste V, Tirode F, Lucchesi C, De Reynies A, Kauffmann A, Bui B, Terrier P, Bonvalot S, Le Cesne A, Vince-

Ranchere D, Blay JY, Collin F, Guillou L, Leroux A, Coindre JM, Aurias A: **Validated prediction of clinical outcome in sarcomas and multiple types of cancer on the basis of a gene expression signature related to genome complexity.** *Nat Med* 2010, **16**:781-787.

8. Korkola JE, Houldsworth J, Feldman DR, Olshen AB, Qin LX, Patil S, Reuter VE, Bosl GJ, Chaganti RS: **Identification and validation of a gene expression signature that predicts outcome in adult men with germ cell tumors.** *J Clin Oncol* 2009, **27**:5240-5247.
9. Lau SK, Boutros PC, Pintilie M, Blackhall FH, Zhu CQ, Strumpf D, Johnston MR, Darling G, Keshavjee S, Waddell TK, Liu N, Lau D, Penn LZ, Shepherd FA, Jurisica I, Der SD, Tsao MS: **Three-gene prognostic classifier for early-stage non small-cell lung cancer.** *J Clin Oncol* 2007, **25**:5562-5569.
10. Van 't Veer LJ, Dai H, Van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, Van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH: **Gene expression profiling predicts clinical outcome of breast cancer.** *Nature* 2002, **415**:530-536.
11. Vermeulen J, De Preter K, Naranjo A, Vercruysse L, Van Roy N, Hellemans J, Swerts K, Bravo S, Scaruffi P, Tonini GP, De Bernardi B, Noguera R, Piqueras M, Canete A, Castel V, Janoueix-Lerosey I, Delattre O, Schleiermacher G, Michon J, Combaret V, Fischer M, Oberthuer A, Ambros PF, Beiske K, Benard J, Marques B, Rubie H, Kohler J, Potschger U, Ladenstein R, *et al*: **Predicting outcomes for children with neuroblastoma using a multigene-expression signature: a retrospective SIOPEN/COG/GPOH study.** *Lnacet Oncol* 2009, **10**:663-671.
12. Zhu CQ, Ding K, Strumpf D, Weir BA, Meyerson M, Pennell N, Thomas RK, Naoki K, Ladd-Acosta C, Liu N, Pintilie M, Der S, Seymour L, Jurisica I, Shepherd FA, Tsao MS: **Prognostic and predictive gene signature for adjuvant chemotherapy in resected non-small-cell lung cancer.** *J Clin Oncol* 2010, **28**:4417-4424.
13. Khambata-Ford S, Garrett CR, Meropol NJ, Basik M, Harbison CT, Wu S, Wong TW, Huang X, Takimoto CH, Godwin AK, Tan BR, Krishnamurthi SS, Burris HA, Poplin EA, Hidalgo M, Baselga J, Clark EA, Mauro DJ: **Expression of epiregulin and amphiregulin and K-ras mutation status predict disease control in metastatic colorectal cancer patients treated with cetuximab.** *J Clin Oncol* 2007, **25**:3230-3237.
14. Quon G, Morris Q: **ISOLATE: a computational strategy for identifying the primary origin of cancers using high-throughput sequencing.** *Bioinformatics* 2009, **25**:2882-2889.
15. Varadhachary GR, Talantov D, Raber MN, Meng C, Hess KR, Jatkoe T, Lenzi R, Spigel DR, Wang Y, Greco FA, Abbruzzese JL, Hainsworth JD: **Molecular profiling of carcinoma of unknown primary and correlation with clinical evaluation.** *J Clin Oncol* 2008, **26**:4442-4448.
16. Bueno-de-Mesquita JM, Van Harten WH, Retel VP, van't Veer LJ, Van Dam FS, Karsenberg K, Douma KF, Van Tinteren H, Peterse JL, Wesseling J, Wu TS, Atsma D, Rutgers EJ, Brink G, Floore AN, Glas AM, Roumen RM, Bellot FE, Van Krimpen C, Rodenhuis S, Van de Vijver MJ, Linn SC: **Use of 70-gene signature to predict prognosis of patients with node-negative breast cancer: a prospective community-based feasibility study (RASTER).** *Lnacet Oncol* 2007, **8**:1079-1087.
17. Glas AM, Floore A, Delahaye LJ, Witteveen AT, Pover RC, Bakx N, Lahti-Domenici JS, Bruinsma TJ, Warmoes MO, Bernards R, Wessels LF, Van't Veer LJ: **Converting a breast cancer microarray signature into a high-throughput diagnostic test.** *BMC Genomics* 2006, **7**:278.
18. Monzon FA, Lyons-Weiler M, Buturovic LJ, Rigl CT, Henner WD, Sciulli C, Dumur CI, Medeiros F, Anderson GG: **Multicenter validation of a 1,550-gene expression profile for identification of tumor tissue of origin.** *J Clin Oncol* 2009, **27**:2503-2508.
19. **Trial watch: Adaptive BATTLE trial uses biomarkers to guide lung cancer treatment..** *Nat Rev Drug Discov* 2010, **9**:423.
20. Stuart RO, Wachsman W, Berry CC, Wang-Rodriguez J, Wasserman L, Klacansky I, Masys D, Arden K, Goodison S, McClelland M, Wang Y, Sawyers A, Kalcheva I, Tarin D, Mercola D: **In silico dissection of cell-type-associated patterns of gene expression in prostate cancer.** *P Natl Acad Sci USA* 2004, **101**:615-620.
21. Wang Y, Xia XQ, Jia Z, Sawyers A, Yao H, Wang-Rodriquez J, Mercola D, McClelland M: **In silico estimates of tissue components in surgical samples based on expression profiling data.** *Cancer Res* 2010, **70**:6448-6455.
22. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M: **Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses.** *P Natl Acad Sci USA* 2001, **98**:13790-13795.
23. West NP, Dattani M, McShane P, Hutchins G, Grabsch J, Mueller W, Treanor D, Quirke P, Grabsch H: **The proportion of tumour cells is an independent predictor for survival in colorectal cancer patients.** *Brit J Cancer* 2010, **102**:1519-1523.
24. Bachtiary B, Boutros PC, Pintilie M, Shi W, Bastianutto C, Li JH, Schwock J, Zhang W, Penn LZ, Jurisica I, Fyles A, Liu FF: **Gene expression profiling in cervical cancer: an exploration of intratumor heterogeneity.** *Clin Cancer Res* 2006, **12**:5632-5640.
25. Okaty BW, Sugino K, Nelson SB: **A quantitative comparison of cell-type-specific microarray gene expression profiling methods in the mouse brain.** *PLoS One* 2011, **6**:e16493.
26. Venet D, Pecasse F, Maenhaut C, Bersini H: **Separation of samples into their constituents using gene expression data.** *Bioinformatics* 2001, **17(Suppl 1)**:S279-87.
27. Lahdesmaki H, Shmulevich L, Dunmire V, Yli-Harja O, Zhang W: **In silico microdissection of microarray data from heterogeneous cell populations.** *BMC Bioinformatics* 2005, **6**:54.
28. Erkkila T, Lehmusvaara S, Ruusuvuori P, Visakorpi T, Shmulevich I, Lahdesmaki H: **Probabilistic analysis of gene expression measurements from heterogeneous tissues.** *Bioinformatics* 2010, **26**:2571-2577.
29. Ghosh D: **Mixture models for assessing differential expression in complex tissues using microarray data.** *Bioinformatics* 2004, **20**:1663-1669.
30. Shen-Orr SS, Tibshirani R, Khatri P, Bodian DL, Staedtler F, Perry NM, Hastie T, Sarwal MM, Davis MM, Butte AJ: **Cell type-specific gene expression differences in complex tissues.** *Nat Methods* 2010, **7**:287-289.
31. Tolliver D, Tsourakakis C, Subramanian A, Shackney S, Schwartz R: **Robust unmixing of tumor states in array comparative genomic hybridization data.** *Bioinformatics* 2010, **26**:i106-14.
32. Bar-Joseph Z, Siegfried Z, Brandeis M, Brors B, Lu Y, Eils R, Dynlacht BD, Simon I: **Genome-wide transcriptional analysis of the human cell cycle identifies genes differentially regulated in normal and cancer cells.** *P Natl Acad Sci USA* 2008, **105**:955-60.
33. Clarke J, Seo P, Clarke B: **Statistical expression deconvolution from mixed tissue samples.** *Bioinformatics* 2010, **26**:1043-1049.
34. Gosink MM, Petrie HT, Tsinoremas NF: **Electronically subtracting expression patterns from a mixed cell population.** *Bioinformatics* 2007, **23**:3328-3334.
35. Polak E, Ribiere G: **Note sur la convergence de méthodes de directions conjuguées.** *ESAIM-Math Model Num* 1969, **3**:35-43.
36. **ISOpure download site..** [http://morrislab.med.utoronto.ca/software.html].
37. **Clarke deconvolution method download site..** [http://biomed.miami.edu/?p=484&pid=185&m=facultyph&mid=1&item=328].
38. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP: **Summaries of Affymetrix GeneChip probe level data.** *Nucleic Acids Res* 2003, **31**:e15.
39. Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H, Watson SJ, Meng F: **Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data.** *Nucleic Acids Res* 2005, **33**:e175.
40. Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, Lizyness ML, Kuick R, Hayasaka S, Taylor JM, Iannettoni MD, Orringer MB, Hanash S: **Gene-expression profiles predict survival of patients with lung adenocarcinoma.** *Nat Med* 2002, **8**:816-824.
41. Shedden K, Taylor JM, Enkemann SA, Tsao MS, Yeatman TJ, Gerald WL, Eschrich S, Jurisica I, Giordano TJ, Misek DE, Chang AC, Zhu CQ, Strumpf D, Hanash S, Shepherd FA, Ding K, Seymour L, Naoki K, Pennell N, Weir B, Verhaak R, Ladd-Acosta C, Golub T, Gruidl M, Sharma A, Szoke J, Zakowski M, Rusch V, Kris M, Viale A, *et al*: **Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study.** *Nat Med* 2008, **14**:822-827.
42. Landi MT, Dracheva T, Rotunno M, Figueroa JD, Liu H, Dasgupta A, Mann FE, Fukuoka J, Hames M, Bergen AW, Murphy SE, Yang P, Pesatori AC, Consonni D, Bertazzi PA, Wacholder S, Shih JH, Caporaso NE, Jen J: **Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival.** *PLoS One* 2008, **3**:e1651.
43. Wallace TA, Prueitt RL, Yi M, Howe TM, Gillespie JW, Yfantis HG, Stephens RM, Caporaso NE, Loffredo CA, Ambs S: **Tumor**

immunobiological differences in prostate cancer between African-American and European-American men. *Cancer Res* 2008, **68**:927-936.
44. Friedman J, Hastie T, Tibshirani R: **Regularization Paths for Generalized Linear Models via Coordinate Descent.** *J Stat Softw* 2010, **33**:1-22.
45. Stephenson AJ, Scardino PT, Eastham JA, Bianco FJ Jr, Dotan ZA, DiBlasio CJ, Reuther A, Klein EA, Kattan MW: **Postoperative nomogram predicting the 10-year probability of prostate cancer recurrence after radical prostatectomy.** *J Clin Oncol* 2005, **23**:7005-7012.
46. Subramanian J, Simon R: **Gene expression-based prognostic signatures in lung cancer: ready for clinical use?.** *J Natl Cancer I* 2010, **102**:464-474.
47. Thompson IM, Tangen CM, Paradelo J, Lucia MS, Miller G, Troyer D, Messing E, Forman J, Chin J, Swanson G, Canby-Hagino E, Crawford ED: **Adjuvant radiotherapy for pathological T3N0M0 prostate cancer significantly reduces risk of metastases and improves survival: long-term followup of a randomized clinical trial.** *J Urology* 2009, **181**:956-962.
48. Magi-Galluzzi C, Evans AJ, Delahunt B, Epstein JI, Griffiths DF, Van der Kwast TH, Montironi R, Wheeler TM, Srigley JR, Egevad LL, Humphrey PA: **International Society of Urological Pathology (ISUP) Consensus Conference on Handling and Staging of Radical Prostatectomy Specimens. Working group 3: extraprostatic extension, lymphovascular invasion and locally advanced disease.** *Modern Pathol* 2011, **24**:26-38.
49. Montanaro L, Trere D, Derenzini M: **Nucleolus, ribosomes, and cancer.** *Am J Pathol* 2008, **173**:301-310.
50. De Marzo AM, Platz EA, Sutcliffe S, Xu J, Grönberg H, Drake CG, Nakai Y, Isaacs WB, Nelson WG: **Inflammation in prostate carcinogenesis.** *Nat Rev Cancer* 2007, **7**:256-69.
51. Nicolau M, Tibshirani R, Børresen-Dale A-L, Jeffrey SS: **Disease-specific genomic analysis: identifying the signature of pathologic biology.** *Bioinformatics* 2007, **23**:957-65.
52. **Integrated genomic analyses of ovarian carcinoma..** *Nature* 2011, **474**:609-15.
53. Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, Gräf S, Ha G, Haffari G, Bashashati A, Russell R, McKinney S, Langerød A, Green A, Provenzano E, Wishart G, Pinder S, Watson P, Markowetz F, Murphy L, Ellis I, Purushotham A, Børresen-Dale A-L, Brenton JD, Tavaré S, Caldas C, *et al*: **The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.** *Nature* 2012, **486**:346-52.
54. Nowak F, Soria J-C, Calvo F: **Tumour molecular profiling for deciding therapy-the French initiative.** *Nat Rev Clin Oncol* 2012, **9**:479-86.
55. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5**:621-8.
56. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.** *Nat Biotechnol* 2010, **28**:511-515.
57. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ: **Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing.** *Nat Genet* 2008, **40**:1413-5.

doi:10.1186/gm433
Cite this article as: Quon *et al.*: Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. *Genome Medicine* 2013 **5**:29.